2010-08-15

Proceedings of the Workshop on Computational Models of Spatial Language Interpretation at Spatial Cognition 2010 (COSLI-2010).

Robert J. Ross
Technological University Dublin, robert.ross@tudublin.ie

Joana Hois
Technological University Dublin, joana@informatik.uni-bremen.de

John D. Kelleher
Technological University Dublin, john.d.kelleher@tudublin.ie

CoSLI 2010
Computational Models of Spatial Language Interpretation
– Preface –

Robert Ross[1], Joana Hois[2], and John Kelleher[1]

[1] Artificial Intelligence Group
Dublin Institute of Technology, Ireland
robert.ross@dit.ie,johnd.kelleher@dit.ie
[2] Research Center on Spatial Cognition (SFB/TR 8)
University of Bremen, Germany
joana@informatik.uni-bremen.de

Computational Models of Spatial Language Interpretation

Competence in spatial language modelling is a cardinal issue in disciplines including Cognitive Psychology, Computational Linguistics, and Computer Science. Within Cognitive Psychology, the relation of spatial language to models of spatial representation and reasoning is considered essential to the development of more complete models of psycholinguistic and cognitive linguistic theories. Within Computer Science and Computational Linguistics, the development of a wide class of so-called situated systems such as robotics, virtual characters, and Geographic Information Systems is heavily dependent on the existence of adequate models of spatial language in order to allow users to interact with these systems when standard graphical, textual, or tactile modes of communication are infeasible or inconvenient.

Competence in spatial language requires that we assign appropriate meaning to spatial terms such as projective, perspective, topological, distance, and path descriptive markers. However, it is not the case that a given linguistic unit such as a spatial preposition has a meaning that can be described in terms of a single qualitative or quantitative model. The same preposition can have multiple meanings, and such variance must be handled through either underspecified models that can be stretched to particular situations, or models which incorporate multiple disparate meanings that are assigned to terms as a situation invites, or models that take into account vague interpretations in situated contexts. In spite of some formal proposals in this area, such heterogeneous meaning accounts are rarely seen in practical computational systems. Moreover, while early models of spatial term interpretation focused on the geometric interpretation of spatial language, it is now widely recognized that spatial term meaning is also dependent on functional and pragmatic features. Competent models of spatial language must thus draw on complex models of situated meaning, and while some early proposals exist, it is not at all clear how geometric, functional and pragmatic features should be integrated in computational models of spatial language interpretation.

The aim of the CoSLI 2010 workshop is to draw together the often orthogonal views on formal symbolic and embodied spatial language interpretation in order to foster theories which adequately draw on both geometric and functional spatial language meaning. On one hand, formal symbolic approaches have attempted to assign meaning to spatial terms through well defined theories that provide a natural symbolic backbone to connect spatial meaning with heterogeneous sources of knowledge and reasoning. These symbolic models, however, often simplify and generalize spatial term meanings and ignore their various situated interpretations. On the other hand, embodied quantitative interpretation models assign meaning to spatial terms through spatial templates which relate the symbolic level to sub-symbolic knowledge such as sensory-motor information and spatial representations more suited to real situated systems. These quantitative models, however, often define templates in a rigid way that allows only few generalizations. By drawing together these formal symbolic and embodied models of spatial meaning we wish to move the research community towards models of spatial meaning which couple embodied geometric and functional features in order to improve and support situated natural language interpretation systems.

Workshop Organization

Organising Committee

Robert Ross	Artificial Intelligence Group, Dublin Institute of Technology, Ireland
Joana Hois	Research Center on Spatial Cognition (SFB/TR 8), University of Bremen, Germany
John Kelleher	Artificial Intelligence Group, Dublin Institute of Technology, Ireland

Programme Committee

John Bateman	University of Bremen, Germany
Brandon Bennett	University of Leeds, UK
Kenny Coventry	Northumbria University, UK
Max J. Egenhofer	University of Maine, USA
Carola Eschenbach	University of Hamburg, Germany
Ben Kuipers	University of Michigan, USA
Reinhard Moratz	University of Maine, USA
Philippe Muller	Université Paul Sabatier, France
Robert Porzel	University of Bremen, Germany
Terry Regier	UC Berkeley, USA
David Schlangen	University of Potsdam, Germany
Andrea Tyler	Georgetown University, Washington, DC, USA

Invited Speaker

Terry Regier	Linguistics and Cognitive Science, UC Berkeley, USA

Acknowledgements

We acknowledge generous financial support from the DFG-funded Research Center on Spatial Cognition (SFB/TR 8) situated at the Universities of Bremen & Freiburg, Germany, and from the Artificial Intelligence Group situated at the Dublin Institute of Technology, Ireland. We would like to thank the PC members for their timely reviewing work and our invited speaker, Terry Regier, for delivering the keynote presentation at the workshop.

We would also like to thank the organizers of the Spatial Cognition 2010 conference for hosting the COSLI workshop, in particular, Adrienne Larmett, Dominique Dumay, Thomas F. Shipley, and Thomas Barkowsky for their support.

August 2010

R. Ross, J. Hois, J. Kelleher
CoSLI 2010 Program Chair

Workshop Schedule

10:30-12:00	Arrival (poster set up)
12:00-12:20	Coffee Break & Informal Start
12:20-12:30	Opening
12:30-13:30	Invited Talk: Terry Regier "Universals and variation in spatial language and thought"
13:30-14:00	Jamie Frost, Alastair Harrison, Stephen Pulman, and Paul Newman "A Probabilistic Approach to Modelling Spatial Language with Its Application To Sensor Models"
14:00-15:00	Lunch (poster set up cont.)
15:00-15:30	Bob Coyne, Richard Sproat, and Julia Hirschberg "Spatial Relations in Text-to-Scene Conversion"
15:30-16:00	Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens "From Language towards Formal Spatial Calculi"
16:00-16:30	Simon Dobnik and Stephen G. Pulman "Human evaluation of robot-generated spatial descriptions"
16:30-17:00	Coffee
17:00-17:30	Poster Session
17:30-18:00	Michael Speriosu, Travis Brown, Taesun Moon, Jason Baldridge, and Katrin Erk "Connecting Language and Geography with Region-Topic Models"
18:00-18:30	Closing
18:30-19:30	Student reception
18:30-19:30	Cash bar
19:30	Conference Opening Dinner

Table of Contents

A Probabilistic Approach to Modelling Spatial Language with Its Application To Sensor Models

Jamie Frost[1], Alastair Harrison[2], Stephen Pulman[1], and Paul Newman[2]

[1] University of Oxford, Computational Linguistics Group, OX1 3QD, UK
{jamie.frost,sgp}@clg.ox.ac.uk
[2] University of Oxford, Mobile Robots Group, OX1 3PJ, UK
{arh,pnewman}@robots.ox.ac.uk

Abstract. We examine why a probabilistic approach to modelling the various components of spatial language is the most practical for spatial algorithms in which they can be employed, and examine such models for prepositions such as 'between' and 'by'. We provide an example of such a probabilistic treatment by exploring a novel application of spatial models to the induction of the occupancy of an object in space given a description about it.

1 Introduction

Space occupies a privileged place in language and our cognitive systems, given the necessity to conceptualise various semantic domains. Spatial language can broadly be divided into two categories [1]: functions which map regions to some part of it, e.g. 'the corner of the park', and functions (in the form of spatial prepositions) which map a region to either an adjacent region, projection or axis, e.g. 'the car between the two trees'. Approaches to implementing spatial models have fallen into two categories. [2] for example takes a logic-based approach, using a set of predicates on objects and binary or tertiary relations that connect objects to generate descriptions of objects that distinguishes it from others. A second approach is a numerical one, which given some reference object or objects and another 'located' object[1] or point, assigns a value based on some notion of 'satisfaction' of the spatial relation in question. But conceptualisation of this assigned value has a large amount of variety. [3] uses a 'Potential Field Model' characterised by potential fields which decreases away from object boundaries. [4] for example uses a linear function to model topological prepositions such as 'near', and produces a value in the range [0,1] depending on whether some point is directly by the object in question or on/beyond the horizon.

However, we argue that a conceptually more rigorous probabilistic approach is needed for all aspects of spatial language, in which validity of some spatial or semantic proposition is determined by the likelihood a human within the context

[1] We use the term 'locative expression' to refer to any expression whose intention is to identify the location of an object or objects (such as 'a chair by the table'). The 'located object' refers to the object in question, and the 'reference' object(s) are others that can be used to determine the location of the located object (the *table* in the latter example).

of the expression would deem it to be true. We motivate this by the following reasons:

1. It provides a uniform treatment of confidence across both spatial and non-spatial domains; uncertainty may be established in the latter in cases of variants of descriptive attributes (such as names) for example. As a result these models can be used in a variety of spatial algorithms such as searching or describing objects and inferring the occupancy in space of an object.
2. In the latter of the above applications (which will be explored in detail) as well as other independent systems or frameworks, a probabilistic representation is often required.
3. Combining multiple spatial observations becomes more transparent: While any monotonically increasing or decreasing function is sufficient to establish a relative measure of applicability across candidate points or objects, the lack of consideration of the function's 'absolute' value becomes problematic when combining data from different spatial models, for example if we were to say 'The chair is by the table *and* between the cat and the rug'.

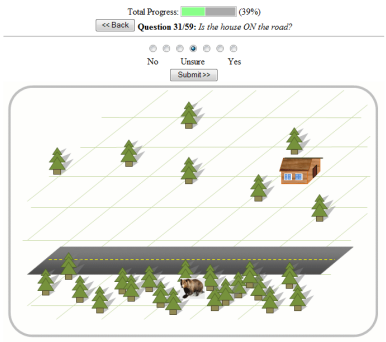

Fig. 1. One of the questions for the 'on/next to/by' section in an online experiment. There were 132 questions in total across the 3 sections.

Such an approach of assessing the 'acceptibility' of regions given a spatial relation is based on a concept called 'Spatial Templates' established by [5], but a probabilistic approach puts more emphasis on *absolute* value. What precisely then do we mean by 'human confidence'? One might think we can measure it by the probability that a given human would consider a (spatial) proposition to be true. But such a notion neglects a concept in philosophy known as *subjectivism*, in which rational agents can have degrees-of-belief in a proposition (rather than constricted to boolean answers of 'agree' or 'disagree'), and probabilities can be interpreted as the measure of such a belief. With such an assumption it is therefore sufficient to construct our models based on the 'average degree-of-belief' across people in some sample. Generically, this confidence can be defined as $p(\phi|\psi)$, where ϕ represents the proposition and ψ represents the context. For a particular spatial model, one might use $p(in\_front(obj_1, obj_2)|x_t)$, where x_t is the current position of the observer. We use $\phi_{\mathbf{x}}$ as a convenience to indicate that the location (say its centre of mass) of the *located object* in ϕ is at position $\mathbf{x}$.

In the next section we present such models we have developed for the prepositions 'between' and 'by', and present a possible novel approach in which we

might induce the occupancy of an object in space given a spatial description. We carried out an online experiment in which users asserted the validity of various locative expressions given a variety of scenes. For each category of spatial relation, e.g. *by* and *between* (and a number of other prepositions not presented here), the user was asked to rate the extent to which they agreed with the given statement, on a scale of 1 (representing 'no') to 7 (representing 'yes'), each question accompanied by a picture[2]. To produce the 'average degree-of-belief' we

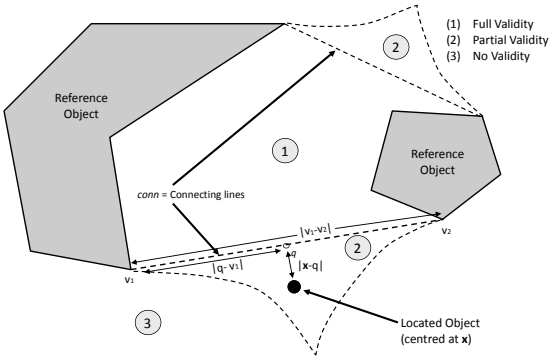

Fig. 2. The variation of confidence for the preposition 'between'.

scaled the average answer to [0,1]. Our models are based on the *Proximal Model* as described in [7]. That is, features are based on the nearest point to the reference object, thus incorporating the shape of the object. This is in contrast to the *Centre-of-Mass Model* (as used in [4] for example) which treats all objects as points. This latter approach is computationally simpler and requires less data, although can be problematic for larger objects; if for example we were to assess the acceptibility of 'you are near the park', we would expect such a judgement to be based on proximity to the edge of the park rather than the centre.

2 Spatial Models for 'between' and 'by'

2.1 "Between"

The model we present below determines the acceptibility of a proposition $\phi = between(a, b, c)$ such that a is the located object, b and c the reference objects, and the position of a is at $\mathbf{x}$. We determined that any point within the convex hull of the two reference objects (excluding the area of the objects themselves) was deemed to be fully valid. Outside of this area, certainty degraded proportional to the centrality of the object. Our model below quantifies these findings:

$$p(\phi_{\mathbf{x}}|x_t) = p(\phi_{\mathbf{x}}) = \begin{cases} max(0, 1 - \frac{|\mathbf{x}-q|}{tol}) \text{ if } \mathbf{x} \notin Hull(ref_1 \cup ref_2) \\ 1 \text{ otherwise} \end{cases} \quad (1)$$

[2] The experiment was restricted to native English speakers only, due to cross-linguistic variations in spatial coding, such as a lack of distinction between different frames of reference (that is, distinguishing between say the deictic interpretation of "in front of the tree" based on the position of the observer, and the intrinsic interpretation based on the salient side of an object, as in "in front of the shop") [6].

$$\text{s.t.} \quad q = \arg\min_{q'}\{|\mathbf{x} - q'||q' \text{ on line } l\}, \quad tol = |v_1 - v_2|k_1(\frac{|q - v_1|}{|v_1 - v_2|})^{k_2}$$
$$l : (v_1, v_2) = \arg\min_{l'}\{|\mathbf{x} - q'||q' \text{ on line } l', l' \in conn\}$$
$$conn = \{(\bar{v}_1, \bar{v}_2)|\bar{v}_1 \in ref_1, \bar{v}_2 \in ref_2, (\bar{v}_1, \bar{v}_2) \in edges(Hull(ref_1 \cup ref_2))\}$$

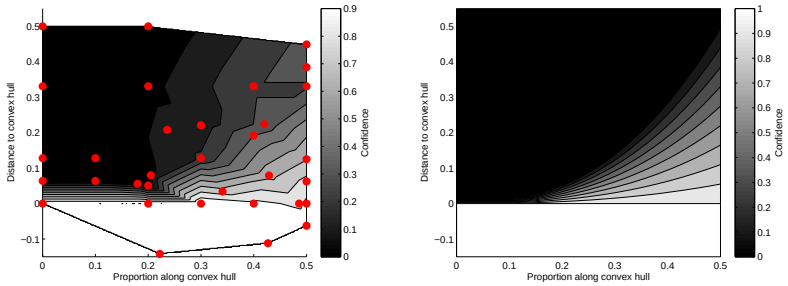

Fig. 3. A comparison of experimental results against the inferred model for 'between'. Both the x and y axis are in terms of the length of the convex hull edge l.

$\mathbf{x}$ is the central point of the located object in question, ref_1 and ref_2 are the vertices of the two referenced objects, $Hull(V)$ gives the convex hull of the set of vertices V (thus q is the nearest point on the convex hull to $\mathbf{x}$), tol gives the maximum allowed distance from the convex hull before the confidence score is 0, $conn$ is the set of 2 edges on the convex hull which connect the shapes corresponding to ref_1 and ref_2 (that is, the straight dotted lines in Fig. 2 and function $edges$ gives the edges of a polygon. k_1 controls the maximum tolerance permitted, a specified proportion of the distance between the two objects, and k_2 controls the curvature of this ambiguous region. Via model fitting (using the minimum sum of squared differences) we found values of $k_1 = 0.55$ and $k_2 = 2.5$ yielded the best results (see Figure 3).

2.2 "By"

For the preposition 'by', there are 3 main variables that can influence the magnitude of the confidence score; the base width (w) and height (h) of the reference object, and the distance (d) from the reference object. For polygonal objects, users were given 8 different reference objects in their scenarios, of a variety of different widths and heights. It was found that although the confidence score for a given distance with respect to the width of the object (i.e. $\frac{d}{w}$) was a good starting point (see Fig 4(a)), greater heights led to a small increase in probability. Assuming a linear relationship with height (again relative to the object width), we therefore divide by $\frac{h}{w} + k_h$ for some constant k_h (given that flat objects such as lakes still yield a non-zero confidence score). Additionally, smaller objects tended to have a larger tolerance of distance with respect to this width, although this effect became less prominent as the width of the object became

very large. Thus we multiply the distance by $\log(w + k_w)$ for some constant $k_w \geq 1$, since for very small objects we still expect some tolerance of distance. Combining these relationships and simplifying, we suggest the following model:

$$p(\phi_{i,j}|x_t) = clamp(k_c - k_m d \frac{\log(w + k_w)}{h + k_h w}) \quad (2)$$

where k_m and k_c the coefficients of some line to obtain the confidence from the adjusted distance, and *clamp* clamps the overall value to the range $[0, 1]$. Fig. 4(b) shows the effect of these using these transformations, using $k_w = 14$ and $k_h = 2$, resulting in values for k_m and k_c of 1.38 and 1.15 respectively. Ultimately it is impossible to base any model of 'by' on physical metrics alone; the 'use case' of objects, i.e. the set of contexts in which an object is used, is likely to have an effect. In Fig. 4(b) for the case where the reference object was a chair, it is apparent confidence deteriorated with distance much faster than expected. But if one considers that a chair is intended 'to be sat on', and therefore adjust the recorded height h to the more salient 'seat-level', we obtain confidence values very close to the model for this example.

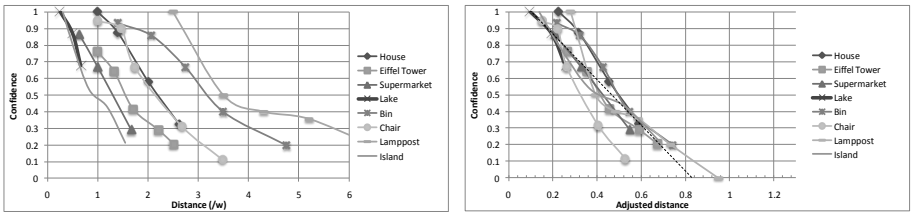

Fig. 4. Experimental results for the preposition 'by' for a number of different examples. Each series indicates the *reference object* used in the locative expression, e.g. 'lake' in "The house is by the lake". Graph (a) shows distance (in terms of width) plotted against validity. Graph (b) shows adjustments as per equation 2.

3 Occupancy Grid Maps

We now propose a method to infer the occupancy in space of a particular object given an observation in the form of a spatial description made about it. This is strongly predicated on a probabilistic treatment of our spatial models discussed earlier. Occupancy Grid Mapping is a technique employed in robotics to generate maps of an environment via noisy sensor measurements. The occupancy grid map is useful because it can subsequently be fused with other maps obtained from say physical sensors. The aim is to produce a posterior $p(m|z_{1:t}, x_{1:t})$ where $m = \{m_i\}$ is a partitioning of space into a finite grid of cells m_i, $z_{1:t}$ are the observations made up to time t, and $x_{1:t}$ are the poses of the robot at each observation. m_i is the event that cell i is occupied, thus $p(m_i)$ describes the probability that cell i is occupied. In the scope of this paper, we focus on how the 'inverse sensor model' $p(m_i|z_t, x_t)$ can be computed, although a more detailed description of Occupancy Grid Mapping can be found in [8].

Our aim is to compute this inverse sensor model, in terms of our calculated $p(\phi|x_t)$ probabilities from the previous section. An important simplifying assumption we make is that locative expressions refer to a *specific point*

in space, within the boundaries of the object in question. This seems intuitive; were we to describe a town as being '10km away', it would clearly be fallacious to assume that the entirety of the town is precisely 10km away. We define a probability $p(r_{i,j}|x_t, z_t)$, where $r_{i,j}$ represents the event that the observer was referring to a point (i, j) in their observation, and z_t is the locative expression such that the position of the located object is not specified, say $\phi_\ominus$ (since we consider such a position in $r_{i,j}$). We can then calculate the desired probability easily by simply normalising our confidence function across the space: $p(r_{i,j}|x_t, z_t = \phi_\ominus) = \alpha p(\phi_{i,j}|x_t)$. Before we determine how to calculate $p(m_{i,j}|x_t, z_t)$, we analyse the conceptual parallelism between traditional Occupancy Grid Mapping and that employed in our linguistic context.

3.1 A Comparison of Sensor Models

On a cursory inspection there are some initial clear similarities that can be drawn between the traditional occupancy grid map and our linguistic variant. Both involve the pose of some observer x_t (although depending on the spatial model this is sometimes irrelevant) and some manifestation of an observation z_t; a physical sensor reading with respect to the traditional approach and a locative expression for the linguistic approach. Upon closer analysis more similarities can be drawn. With a physical sensor, we expect a measurement of distance to a point being sensed to be noisy, and thus maintain a probability distribution with regards to the precise position of this point. This corresponds to our distribution $p(r_{i,j}|x_t, z_t)$. For a locative expression of a town being "10km away", human error or rounding is likely to lead to uncertainty in the judged distance, and additionally the direction of the town is unspecified, leading to a 'blurred doughnut' type distribution.

There are however a number of conceptual differences. With traditional Occupancy Grid Mapping the posterior for a cell is only updated if it was part of the sensor range (i.e. we make no assumption with regards to space outside the limited range of our sensor). With locative expressions however, we can infer data outside that explicitly conveyed. Suppose for our town example, the town was 1km in diameter, and that the distance judgement of 10km (to some point within the town) was entirely accurate. If the centre of the town was actually 10.5km away, our observation would still hold, but a point any further could not possibly be occupied by 'town'.

3.2 Computing the Inverse Sensor Model

We can use the above fact to compute $p(m_{i,j}|x_t, z_t)$ from our previously calculated values of $p(r_{i,j}|x_t, z_t)$. Let $\mathcal{Q}$ be the set of possible 'poses' for the located object such that the point (i, j) is within the object's boundary, and a pose is the position and orientation of the object. Given our assumption that the observer referred to a point within the confines of their perceived position of the object, $\mathcal{Q}$ represents all valid poses of the object given such a point. It follows that $p(m_{i,j}|x_t, z_t) = \int_{q\in\mathcal{Q}} p(q|x_t, z_t)\,\mathrm{d}q$. For each pose $q \in \mathcal{Q}$ there is an associated frame (i^*_q, j^*_q, θ_q) where (i^*_q, j^*_q) is the *nominal centre* of the shape in pose q (say

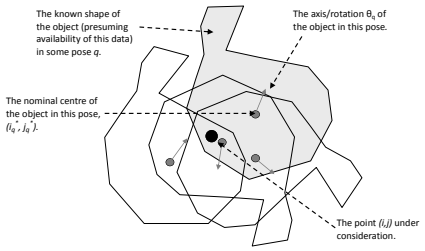

Fig. 5. In calculating the occupancy probability $p(m_{i,j}|x_t, z_t)$, we consider all possible poses of the located object in which the point (i, j) is confined. The probability of each pose q is $p(r_{i_q^*, j_q^*}|x_t, z_t)p(\theta_q)$.

the centre of mass) and θ_q is the rotation of the object about this point. It is then possible to use $p(r_{i_q^*, j_q^*}|x_t, z_t)$ to refer to the probability of the object being positioned at (i_q^*, j_q^*) (see Fig. 5). The pose also has a probability $p(\theta_q)$ associated with its orientation; for simplicity we assume this is independent of x_t and z_t (although the use of $p(\theta_q|z_t)$ would allow us to model for example observations such as "The boat is in front of you, *facing East*"). Putting this together, this gives us the following equation to compute the occupancy probability:

$$p(m_{i,j}|x_t, z_t) = \int_{q \in \mathcal{Q}} p(r_{i_q^*, j_q^*}|x_t, z_t)p(\theta_q)\,\mathrm{d}q \quad \text{s.t. } \mathcal{Q} = \{q \,|\, (i, j) \in R(q)\} \tag{3}$$

where $R(q)$ is the region of the located object in pose q. Considering the pose of the located object has useful consequences; it allows us to model for example that vehicles are aligned to the direction of a road. Given a lack of prior shape information with regards to the located object, and given the above integral is somewhat intractable, a suitable approximation is to use the approximate width W of the object (which can be obtained via knowledge of the class of the located object, say the usual width of a town). If we infer as little about the shape as possible, the resulting approximation of the shape is a circle of diameter W. Equation 3 then reduces to the following:

$$p(m_{i,j}|x_t, z_t) = \int_{(i',j') \in R(\frac{W}{2}, i, j)} p(r_{i',j'}|x_t, z_t)\,\mathrm{d}i'\,\mathrm{d}j' \tag{4}$$

s.t. $R(\frac{W}{2}, i, j)$ is a set of points in a circular region of centre (i, j) and radius $\frac{W}{2}$

4 Conclusions & Future Work

In this paper we motivated a probabilistic approach to modelling spatial language that can be used in a number of algorithms, and provided an example of such an algorithm to induce a sense of 'the space that an object occupies' via the use of occupancy grid maps. We also presented models for the prepositions

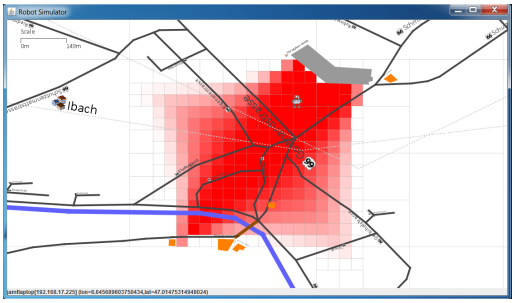

Fig. 6. The occupancy grid generated by a 'between' observation for two objects in an environment. Note that the grid consists of cells of variable size; such a modification to the OGM allows us to choose a cell size appropriate to the scale of the observation, as well as represent areas of constant probability or empty space efficiently.

'by' and 'between' based on the results of an online experiment. Future work is predominantly focused further development of our dialogue manager language that interacts with these spatial models, as well as developing further algorithms which make use of such models. For example, we developed an algorithm that combines semantic and spatial models to provide confidence scores for arbitrarily complex locative expressions (including those based on current bounded trajectories, such as 'the second left'). We are also investigating a measure of 'relevance' (one of the Gricean maxims [9]) for locative expressions, a consideration that is particularly key in generating descriptions of objects or locations.

This work has been supported by the European Commission under grant agreement number FP7-231888-EUROPA.

References

1. Herskovits, A. (1986) Language and Spatial Cognition, Cambridge University Press, .
2. Dale, R. and Haddock, N. (1991) In Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics Morristown, NJ, USA: Association for Computational Linguistics. pp. 161–166.
3. Olivier, P. and Tsujii, J.-I. (2004) *Artificial Intelligence Review* **8**, 147–158.
4. Kelleher, J. D. and Costello, F. J. (2009) *Comput. Linguist.* **35(2)**, 271–306.
5. Logan, G. and Sadler, D. Language and space chapter A computational analysis of the apprehension of spatial relations, pp. 493–529 MIT Press (1996).
6. Levinson, S. C. and Wilkins, D. P. Grammars of Space: Explorations in Cognitive Diversity chapter 1, pp. 4–5 Cambridge University Press (2006).
7. Regier, T. and Carlson, L. A. (2001) *J. Exp Psychol Gen* **130(2)**, 273–298.
8. Thrun, S. (2003) *Auton. Robots* **15(2)**, 111–127.
9. Grice, H. P. (1975) Logic and conversation In Peter Cole and Jerry L. Morgan, (ed.), Syntax and semantics 3: Speech Acts, volume **3**, pp. 41–58 New York: Academic Press.

Spatial Relations in Text-to-Scene Conversion

Bob Coyne[1], Richard Sproat[2], and Julia Hirschberg[1]

[1] Columbia University, New York NY, USA,
{coyne, julia}@cs.columbia.edu,
[2] Oregon Health & Science University, Beaverton, Oregon, USA
rws@xoba.org

Abstract. Spatial relations play an important role in our understanding of language. In particular, they are a crucial component in descriptions of scenes in the world. WordsEye (www.wordseye.com) is a system for automatically converting natural language text into 3D scenes representing the meaning of that text. Natural language offers an interface to scene generation that is intuitive and immediately approachable by anyone, without any special skill or training. WordsEye has been used by several thousand users on the web to create approximately 15,000 fully rendered scenes. We describe how the system incorporates geometric and semantic knowledge about objects and their parts and the spatial relations that hold among these in order to depict spatial relations in 3D scenes.

1 Introduction

Spatial relations are expressed either directly or implicitly in a wide range of natural language descriptions. To represent these descriptions in a 3D scene, one needs both linguistic and real-world knowledge, in particular knowledge about: the spatial and functional properties of objects; prepositions and the spatial relations they convey, which is often ambiguous; verbs and how they resolve to poses and other spatial relations. For example, to interpret *apple in the bowl* we use our knowledge of bowls – that they have interiors that can contain objects. With different objects (e.g., *boat in water*), a different spatial relation is conveyed.

WordsEye [6] is a system for automatically converting natural language text into 3D scenes representing the meaning of that text. A version of WordsEye has been tested online (www.wordseye.com) with several thousand real-world users. We have also performed preliminary testing of the system in schools, as a way to help students exercise their language skills. Students found the software fun to use, an important element in motivating learning. As one teacher reported, "One kid who never likes anything we do had a great time yesterday...was laughing out loud."

WordsEye currently focuses on directly expressed spatial relations and other graphically realizable properties. As a result, users must describe scenes in somewhat stilted language. See Figure 1. Our current research focuses on improving the system's ability to infer these relations automatically. However, in this paper, we describe the basic techniques used by WordsEye to interpret and depict directly expressed spatial relations.

In Section 2 we describe previous systems that convert natural language text to 3D scenes and prior linguistic work on spatial relations. In Section 3 we provide an overview of WordsEye. In Section 4 we discuss the spatial, semantic and functional knowledge about objects used to depict spatial relations in our system. We conclude and describe other ongoing and future work in Section 5.

2 Prior Work

Natural language input has been investigated in some early 3D graphics systems [1][13] including the Put system [4], which was limited to spatial arrangements of existing objects in a pre-constructed environment. In this system, input was restricted to an artificial subset of English consisting of expressions of the form $Put(X, P, Y)$, where X and Y are objects and P is a rigidly defined spatial preposition. Work at the University of Pennsylvania's Center of Human Modeling and Simulation [2], used language to control animated characters in a closed virtual environment. CarSim [7] is a domain-specific system that creates animations from natural language descriptions of accident reports. CONFUCIUS [12] is a multi-modal text-to-animation system that generates animations of virtual humans from single sentences containing an action verb. In these systems the referenced objects, attributes, and actions are typically relatively small in number or targeted to specific pre-existing domains.

Spatial relations have been studied in linguistics for many years. One reasonably thorough study for English is Herskovits [9], who catalogs fine-grained distinctions in the interpretations of various prepositions.[3] For example, she distinguishes among the various uses of *on* to mean "on the top of a horizontal surface" (*the cup is on the table*), or "affixed to a vertical surface" (*the picture is on the wall*). Herskovits notes that the interpretation of spatial expressions may involve considerable inference. For example, the sentence *the gas station is at the freeway* clearly implies more than just that the gas station is located next to the freeway; the gas station must be located on a road that passes over or under the freeway, the implication being that, if one proceeds from a given point along that road, one will reach the freeway, and also find the gas station.

[3] It is important to realize that how spatial relations are expressed, and *what kinds of relations may be expressed* varies substantially across languages. Levinson and colleagues [11] have catalogued profound differences in the ways different languages encode relations between objects in the world. In particular, the Australian language Guugu Yimithirr and the Mayan language Tzeltal use absolute frames of reference to refer to the relative positions of objects. In Guugu Yimithirr, one can locate a chair relative to a table only in terms of cardinal points saying, for example, that the chair is north of the table. In English such expressions are reserved for geographical contexts — *Seattle is north of Portland* — and are never used for relations at what Levinson terms the "domestic scale". In Guugu Yimithirr one has no choice, and there are no direct translations for English expressions such as *the chair is in front of the table*.

Eye of the Beholder by Bob Coyne

Input text: The silver penny is on the moss ground. The penny is 7 feet tall. A clown is 2 feet in front of the penny. The clown is facing the penny.

No Dying Allowed by Richard Sproat

Input text: Eight big white washing machines are in front of the big cream wall. The wall is 100 feet long. The "No Dying Allowed" whiteboard is on the wall. The whiteboard is one foot high and five feet long. The ground is tile. Death is in front of the washing machines. It is facing southeast. Death is eight feet tall.

Fig. 1: Some Examples from WordsEye's Online Gallery

3 System Overview

Our current system is an updated version of the original WordsEye system [6], which was the first system to use a large library of 3D objects to depict scenes in a free-form manner using natural language. The current system contains 2,200 3D objects and 10,000 images and a lexicon of approximately 15,000 nouns. It supports language-based control of objects, spatial relations, and surface properties (e.g., textures and colors); and it handles simple coreference resolution, allowing for a variety of ways of referring to objects. The original WordsEye system handled 200 verbs in an *ad hoc* manner with no systematic semantic modeling of verb alternations and argument combinations. In the current system, we are instead adding frame semantics to support verbs more robustly. To do this, we are utilizing our own lexical knowledge-base, called the SBLR (Scenario-Based Lexical Resource) [5]. The SBLR consists of an ontology and lexical semantic information extracted from WordNet [8] and FrameNet [3] which we are augmenting to include the finer-grained relations and properties on entities needed for depicting scenes as well as capturing the different senses of prepositions related to those properties and relations.

The system works by first parsing each input sentence into a dependency structure. These dependency structures are then processed to resolve anaphora and other coreferences. The lexical items and dependency links are then converted to semantic nodes and roles drawing on lexical valence patterns and other information in the SBLR. The resulting semantic relations are then converted to a final set of graphical constraints representing the position, orientation, size, color, texture, and poses of objects in the scene. Finally, the scene is composed from these constraints and rendered in OpenGL (`http://www.opengl.org`) and

optionally ray-traced in Radiance [10]. The user can then provide a title and caption and save the scene in our online gallery where others can comment and create their own pictures in response. See Figure 1.

4 Spatial Relations

WordsEye uses SPATIAL TAGS and other spatial and functional properties on objects to resolve the meaning of spatial relations. We focus here on the interpretation of NPs containing spatial prepositions of the form "X-preposition-Y", where we will refer to X as the FIGURE and Y as the GROUND. For example, in *snow is on the roof*, *snow* is the FIGURE and *roof* is GROUND. The interpretation of the spatial relation often depends upon the types of the arguments to the preposition. There can be more than one interpretation of a spatial relation for a given preposition The geometric and semantic information associated with those objects will, however, help narrow down the possibilities.

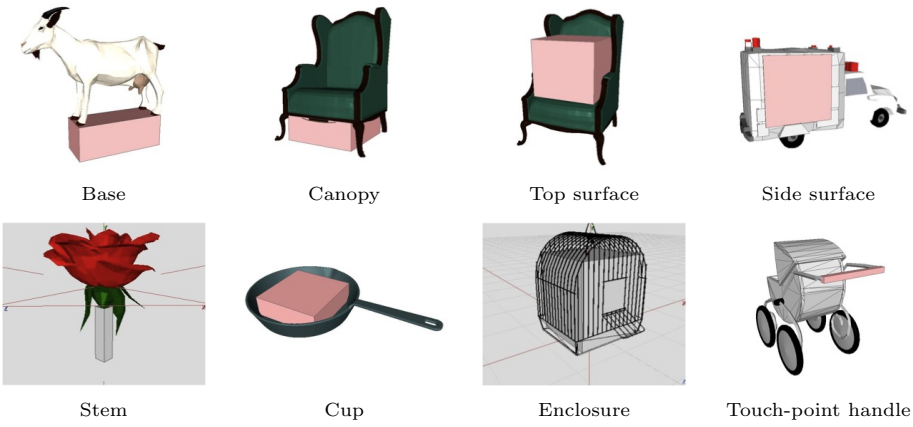

Fig. 2: Spatial Tags, represented here by the boxes associated with each object, designate regions of those objects used in resolving spatial relations. For example, the TOP SURFACE region marked on the seat of the chair is used in sentences like *The pink mouse is on the small chair* to position the FIGURE (*mouse*) on the GROUND (*chair*). See Figure 3 for the depiction of this sentence and several others that illustrate the effect of spatial tags and other object features.

The 3D objects in our system are augmented with the following features:

- IS-A: The lexical category to which the given object belongs.
- Spatial tags identifying the following regions: (See Figure 2)
  - CANOPY: A canopy-like area "under" an object (e.g., *under a tree*).
  - CUP: A hollow area, open above, that forms the interior of an object.
  - ENCLOSURE: An interior region, bounded on all sides (holes allowed).

  - TOP/SIDE/BOTTOM/FRONT/BACK: For both inner and outer surfaces.
  - NAMED-PART: For example, the hood on car.
  - STEM: A long thin, vertical base.
  - OPENING: An opening to an object's interior (e.g., doorway to a room).
  - HOLE-THROUGH: A hole through an object. For example, a ring or donut.
  - TOUCH-POINT: Handles and other functional parts on the object. For example, in *John opened the door*, the doorknob would be marked as a handle, allowing the hand to grasp at that location.
  - BASE: The region of an object where it supports itself.
- OVERALL SHAPE: A dominant overall shape used in resolving various spatial relations. For example, SHEET, BLOCK, RIBBON, CUP, TUBE, DISK, ROD.
- FORWARD/UPRIGHT DIRECTION: The object's default orientation.
- SIZE: The default real-world size of the object. This is also used in spatial relations where the FIGURE and GROUND size must be compatible. For example, *ring on a stick* versus **life-preserver on a pencil*.
- LENGTH AXIS: The axis for lengthening an object.
- SEGMENTED/STRETCHABLE: Some objects don't change size in all dimensions proportionally. For example, a fence can be extended indefinitely in length without a corresponding change in height.
- EMBEDDABLE: Some objects, in their normal function, are embedded in others. For example, fireplaces are embedded in walls, and boats in water.
- WALL-ITEM and CEILING-ITEM: Some objects are commonly attached to walls or ceilings or other non-upward surfaces. Some (e.g., pictures) do this by virtue of their OVERALL SHAPE, while for others (e.g., sconces) the orientation of the object's BASE is used to properly position the object.
- FLEXIBLE: Flexible objects such as cloth and paper allow an object to hang or wrap. For example, *towel over a chair*.
- SURFACE ELEMENT: Any object that can be part of a flat surface or layer. For example, a crack, smudge, decal, or texture.
- Semantic properties such as PATH, SEAT, AIRBORNE for object function.

Some of these features were used in earlier versions of our system [6]. Features we have added to the current version include: SURFACE ELEMENT, EMBEDDABLE, OVERALL SHAPE, LENGTH AXIS, SEGMENTED/STRETCHABLE. Other features, including (FLEXIBLE, OPENING, HOLE-THROUGH and various semantic features) are in the development stage. The implemented tagset supports the generation of scenes such as Figure 3.

In order to resolve a spatial relation, we find the spatial tags and other features of the FIGURE and GROUND objects that are applicable for the given preposition. For example, if the preposition is *under*, a CANOPY region for the GROUND object is relevant, but not a TOP SURFACE. Various other factors, such as size, must also be considered. With ENCLOSED-IN, the FIGURE must fully fit in the GROUND. For EMBEDDED-IN, only part need fit. For other relations (e.g., NEXT-TO), the objects can be any size, but the FIGURE location might vary. For example, *The mosquito is next to the horse* and *The dog is next to the horse* position the FIGURE in different places, either in the air or on the ground,

Spatial Relation	**Example**	**Partial Conditions**
on-top-surface	*vase on table*	GROUND is UPWARD-SURFACE
on-vertical-surface	*postcard on fridge*	GROUND is VERTICAL-SURFACE
on-downward-surface	*fan on ceiling*	GROUND is DOWNWARD-SURFACE
on-outward-surface	*pimple on nose*	GROUND is SURFACE
pattern/coating-on	*plaid pattern on shirt*	FIGURE is TEXTURE or LAYER
fit-on-custom	*train on track*	SPECIAL BASE PAIRING
ring-on-pole	*bracelet on wrist*	FIGURE=RING-SHAPE, GROUND=POLE-SHAPE
on-vehicle	*man on bus*	GROUND=PUBLIC-TRANSPORTATION
on-region	*on the left side of...*	GROUND=REGION-DESIGNATOR
hang-on	*towel on rod*	FIGURE is HANGABLE
embedded-in	*pole in ground*	GROUND is MASS
embedded-in	*boat in water*	FIGURE is EMBEDDABLE
buried-in	*treasure in ground*	GROUND is TERRAIN
enclosed-in-volume	*bird in cage*	GROUND has ENCLOSURE
enclosed-in-area	*tree in yard*	GROUND is AREA
in-2D-representation	*man in the photo*	GROUND is 2D REPRESENTATION
in-cup	*cherries in bowl*	GROUND has CUP
in-horiz-opening	*in doorway*	GROUND has OPENING
stem-in-cup	*flower in vase*	FIGURE has STEM, GROUND has CUP
wrapped-in	*chicken in the foil*	GROUND is FLEXIBLE/SHEET
member-of-arrangement	*plate in stack*	GROUND is ARRANGEMENT
in-mixture	*dust in air*	FIGURE/GROUND=SUBSTANCE
in-entanglement	*bird in tree*	GROUND has ENTANGLEMENT
fitted-in	*hand in glove*	FIGURE/GROUND=FIT
in-grip	*pencil in hand*	GROUND=GRIPPER

Table 1: Spatial relations for *in* and *on* (approximately half are currently implemented). Similar mappings exist for other prepositions such as *under*, *along*. Handcrafted rules resolve the spatial relation given the object features.

depending on whether the given object is commonly airborne or not. We also note that the FIGURE is normally the smaller object while the GROUND functions as a landmark. So it's normal to say *The flower bed is next to the house*, but unnatural to say **The house is next to the flowerbed*. This is discussed in [9]. See Table 1 for some mappings we make from prepositions to spatial relations.

In order to use the object features described above to resolve the spatial meaning of prepositions, linguistically referenced subregions must also be considered. Spatial relations often express regions relative to an object (e.g., *left side of* in *The chair is on the left side of the room*). The same subregion designation can yield different interpretations, depending on the features of the objects.

- EXTERNAL-VERTICAL-SURFACE: *shutters on the left side of the house*
- INTERIOR-VERTICAL-SURFACE: *picture on the left side of the room*
- REGION-OF-HORIZ-SURFACE: *vase on the left side of the room*
- NEIGHBORING-AREA: *car on the left side of the house*

These regions (when present) are combined with the other constraints on spatial relations to form the final interpretation.

Input text: *A large magenta flower is in a small vase. The vase is under an umbrella. The umbrella is on the right side of a table. A picture of a woman is on the left side of a 16 foot long wall. A brick texture is on the wall. The wall is 2 feet behind the table. A small brown horse is in the ground. It is a foot to the left of the table. A red chicken is in a birdcage. The cage is to the right of the table. A huge apple is on the wall. It is to the left of the picture. A large rug is under the table. A small blue chicken is in a large flower cereal bowl. A pink mouse is on a small chair. The chair is 5 inches to the left of the bowl. The bowl is in front of the table. The red chicken is facing the blue chicken. . .*

Fig. 3: Spatial relations and features: ENCLOSED-IN (*chicken in cage*); EMBEDDED-IN (*horse in ground*); IN-CUP (*chicken in bowl*); ON-TOP-SURFACE (*apple on wall*); ON-VERTICAL-SURFACE (*picture on wall*); PATTERN-ON (*brick texture on wall*); UNDER-CANOPY (*vase under umbrella*); UNDER-BASE (*rug under table*); STEM-IN-CUP (*flower in vase*); LATERALLY-RELATED (*wall behind table*); LENGTH-AXIS (*wall*); DEFAULT SIZE/ORIENTATION (all objects); REGION (*right side of*); DISTANCE (*2 feet behind*); SIZE (*small* and *16 foot long*); ORIENTATION (*facing*).

5 Conclusions and Ongoing and Future Work

In order to represent spatial relations more robustly, much remains to be done at the language, graphical, and application levels.

We are augmenting the system to resolve verbs to semantic frames using information in our SBLR, and mapping those in turn to corresponding poses and spatial relations [5]. Figure 4 illustrates this process, which currently is supported for a limited set of verbs and their arguments. This enhanced capability also requires contextual information about actions and locations that we are acquiring using human annotations obtained via Amazon's Mechanical Turk and by extracting information from corpora using automatic methods [14]. We will be evaluating our software in partnership with a non-profit after-school program in New York City.

Acknowledgments

This work was supported in part by the NSF IIS- 0904361. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

The truck chased the man down the road...

The man ran across the sidewalk...

Fig. 4: Spatial relations derived from verbs. The verbs are mapped to semantic frames which in turn are mapped to VIGNETTES (representing basic contextual situations) given a set of semantic role and values. These, in turn, are mapped to spatial relations. In the first example, the PURSUED (*soldier*) is in a running pose, located on the PATH (*road*), and in front of the PURSUER (truck).

References

1. Adorni, G., Di Manzo, M., Giunchiglia, F.: Natural language driven image generation. COLING pp. 495–500 (1984)
2. Badler, N., Bindiganavale, R., Bourne, J., Palmer, M., Shi, J., Schule, W.: A parameterized action representation for virtual human agents. Workshop on Embodied Conversational Characters, Lake Tahoe (1998)
3. Baker, C., Fillmore, C., Lowe, J.: The Berkeley FrameNet Project. COLING-ACL (1998)
4. Clay, S.R., Wilhelms, J.: Put: Language-based interactive manipulation of objects. IEEE Computer Graphics and Applications pp. 31–39 (1996)
5. Coyne, B., Rambow, O., Hirschberg, J., Sproat, R.: Frame semantics in text-to-scene generation. 14th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (2010)
6. Coyne, B., Sproat, R.: WordsEye: An automatic text-to-scene conversion system. SIGGRAPH, Computer Graphics Proceedings pp. 487–496 (2001)
7. Dupuy, S., Egges, A., Legendre, V., Nugues, P.: Generating a 3d simulation of a car accident from a written description in natural language: The carsim system. Proceedings of ACL Workshop on Temporal and Spatial Information Processing pp. 1–8 (2001)
8. Fellbaum, C.: WordNet: an electronic lexical database. MIT Press (1998)
9. Herskovits, A.: Language and Spatial Cognition: an Interdisciplinary Study of the Prepositions in English. Cambridge University Press, Cambridge, England (1986)
10. Larson, G., Shakespeare, R.: Rendering with Radiance. The Morgan Kaufmann Series in Computer Graphics (1998)
11. Levinson, S.: Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge (2003)
12. Ma, M.: Automatic Conversion of Natural Language to 3D Animation. Ph.D. thesis, University of Ulster (2006)
13. Simmons, R.: The clowns microworld. Proceedings of TINLAP pp. 17–19 (1998)
14. Sproat, R.: Inferring the environment in a text-to-scene conversion system. First International Conference on Knowledge Capture, Victoria, BC (2001)

From Language towards Formal Spatial Calculi

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens

Katholieke Universiteit Leuven, Departement Computerwetenschappen
{parisa.kordjamshidi,martijn.vanotterlo,sien.moens}@cs.kuleuven.be

Abstract. We consider mapping unrestricted natural language to formal spatial representations. We describe ongoing work on a two-level machine learning approach. The first level is linguistic, and deals with the extraction of spatial information from natural language sentences, and is called *spatial role labeling*. The second level is ontological in nature, and deals with mapping this linguistic, spatial information to formal spatial calculi. Our main obstacles are the lack of available annotated data for training machine learning algorithms for these tasks, and the difficulty of selecting an appropriate abstraction level for the spatial information. For the linguistic part, we approach the problem in a gradual way. We make use of existing resources such as The Preposition Project (TPP) and the validation data of General Upper Model (GUM) ontology, and we show some computational results. For the ontological part, we describe machine learning challenges and discuss our proposed approach.

1 Introduction

An essential function of language is to convey *spatial relationships* between objects and their relative locations in a space. It is a challenging problem in robotics, navigation, query answering systems, etc. [19]. Our research considers the extraction of spatial information in a multimodal environment. We want to represent spatial information using formal representations that allow spatial reasoning. An example of an interesting multimodal environment is the domain of navigation where we expect a robot to follow navigation instructions. By placing a camera on the robot, it should be able to recognize visible objects and their location. In this context, mapping natural language to a formal spatial representation [4] has several advantages. First, generating language from vision and vice versa visualizing the language is more feasible if a formal intermediate layer is employed [16]. Second, applying the same representation for extraction from image/video data allows combining multimodal features for better recognition and disambiguation in each modality. Finally, a unified representation for various modalities enables spatial reasoning based on multimodal information.
In our work we identify two main layers of information (see also [2]):
1) a **linguistic** layer, in which (unrestricted) natural language is mapped onto ontological structures that convey spatial information, and 2) a **formal** layer, in which the ontological information is mapped onto a specific spatial calculus such as *region connection calculus* (RCC) (cf. [4]). For example, in the sentence the

book is on the table the first step should identify that there is a spatial relation (on) between book and table, after which this could be mapped to a specific, formal relation `AboveExternallyConnected(book, table)` between two tokens `book` and `table` that denote two physical objects in some Euclidean space. For both transformations we propose *machine learning* techniques to deal with the many sources of ambiguity in this task. This has not been done systematically before; most often a restricted language is used to extract highly specific and application-dependent relations and usually one focuses on phrases of which it is known that spatial information is present [8, 6, 19, 11].

To apply machine learning effectively, a clear task definition as well as annotated data are needed. Semantic hand-labeling of natural language is an ambiguous, complex and expensive task and in our two-level view we have to cope with the lack of available data two times. In our recently proposed semantic labeling scheme [10], we tag sentences with the spatial roles according to *holistic spatial semantic* (HSS) theory [21] and also formal spatial relation(s). For mapping between language and spatial information, we defined *spatial role labeling* and performed experiments on the (small amount of) available annotated corpora. The Preposition Project (TPP) data is employed for spatial preposition recognition in the context of learning the main spatial roles *trajector* and *landmark* from data. We have conducted initial experiments on the small corpus of the GUM [1] spatial ontology, and the results indicate that machine learning based on linguistic features can indeed be employed for this task.

The second layer of our methodology consists of mapping the extracted spatial information onto formal spatial systems capable of spatial reasoning. Here we propose to annotate data with spatial calculi relations and use machine learning to obtain a *probabilistic logical* model [3] of spatial relations for this mapping. Such models can deal with both the structural aspects of spatial relations, as well as the intrinsic ambiguity and vagueness in such mappings (see also [5]). In the following sections we will describe both the linguistic and the formal steps, and results of our initial machine learning experiments.

2 Linguistic Level and Spatial Role Labeling

To be able to map natural language to spatial calculi we should first extract the components of spatial information. We call this task **spatial role labeling**. It has not been well-defined before and has not been considered as a stand-alone linguistic task. We define it analogous to *semantic role labeling* (SRL) [15], targeting semantic information associated with specific phrases (usually verbs), but as a stand-alone linguistic task utilizing specific (data) resources.

Task definition. We define spatial role labeling (SpRL) as *the automatic labeling of natural language with a set of spatial roles.* The sentence-level spatial analysis of text deals with characterizing spatial descriptions, denoting the spatial properties of objects and their location (e.g. to answer "what/who/where"-questions). A spatial term (typically a preposition) establishes the type of spatial relation and other constituents express the participants of the spatial relation

(e.g. a location). The *roles* are drawn from a pre-specified list of possible spatial roles and the role-bearing constituents in a spatial expression must be identified and their correct spatial role labels assigned.

Representation based on spatial semantics. The spatial role set we employ contains the core roles of *trajector, landmark, spatial indicator*, and *motion indicator* [6, 21], as well as the features *path* and *frame of reference*. Our set of spatial roles are motivated by the theory of holistic spatial semantics upon which we have defined an annotation scheme in [10]. We describe these terms briefly. A **trajector** is the entity whose (trans)location is of relevance. It can be static or dynamic; a person or an object. It can also be expressed as a whole event. Other terms often used for this concept are the *local object*, *locatum*, *figure object*, *referent* and *target*. A **landmark** is the reference entity in relation to which the location or the trajectory of motion of the trajector is specified. Alternative terms include *reference object*, *ground* and *relatum*. A **spatial indicator** is a token which defines constraints on the spatial properties, such as the location of the trajector with respect to the landmark (e.g. in, on). It explains the type of the spatial relation and usually is a preposition, but can also be a verb, a noun, etc. It is the pivot of a spatial relation, and in terms of GUM ontology it is called a spatial modality. A **motion indicator** is a spatial term which is an indicator of motion, e.g. motion *verbs*. We also consider other conceptual aspects like **frame of reference** and the **path** of a motion that are important for spatial semantics and roles [21].

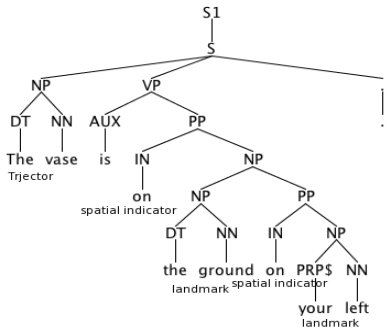

Fig. 1. Parse tree with spatial roles

Linguistic challenges. Given a sentence, SpRL should answer: **Q1.** Does the sentence contain spatial information? **Q2.** What is the **pivot** of the spatial information? (spatial indicator) **Q3.** Starting from the pivot how can we identify/classify the related arguments with respect to predefined set of spatial roles? Spatial relations in English are mostly expressed using prepositions [7], but verbs and even other lexical categories can be central spatial terms. Hence SpRL consists of identifying the boundaries of the arguments of the identified spatial term and then labeling them with spatial roles (argument classification). However, there are very sparse and limited resources for learning spatial roles. Other work typically uses a limited set of words, often based on a set of spatial prepositions and specific grammatical patterns in a specific domain [13, 8].

General extraction of spatial relations is hindered by several things. First, there is not always a regular mapping between a sentence's parse tree and its spatial semantic structure. This is more challenging in complex expressions which convey several spatial relations [4]; see the following sentence (and Fig. 1).

The vase is **on** the ground **on** your left.

Here a dependency parser relates the first "on" to "vase" and "ground". This will produce a valid spatial relation. But the second "on" is related to "ground" and "left", producing a meaningless spatial relation (ground on your left). For more complex relations and nested noun phrases, deriving spatially valid relations is not straightforward and depends on the *lexical* meaning of the words. Other linguistic phenomena such as *spatial-focus-shift* and *ellipsis of trajector and landmark* [11] make the extraction more difficult. Recognizing the right PP-attachment (i.e. whether the preposition is attached to the verb phrase or noun phrase) could help the identification of spatial arguments when the verb in the sentence conveys spatial meaning. *Spatial motion detection* and *recognition of the frame of reference* are additional challenges but will not be dealt with here.
Approach. We aim to tackle the problem using machine learning, in a way similar to SRL, but with important differences. The first difference is that the main focus of SRL is on the *predicate*, its related arguments and their roles [15]. On the other hand, in SpRL the spatial indicator plays the main role and should be identified and disambiguated beforehand. Second, the set of main roles is quite different in SpRL and a large enough English corpus is not available from which spatial roles can be learned directly. Hence new data resources are needed. The main point is that we aim at domain-independent and unrestricted language analysis. This prohibits using very limited data or a small set of extraction rules. However, utilizing existing linguistic resources which can partially or indirectly help to set up a (relational) joint learning framework will be of great advantage. It can relinquish the necessity of expensive labeling of one huge corpus. Our results for preliminary experiments are briefly described in Section 4.

3 Towards Spatial Calculi and Spatial Formalizing

Mapping the spatial information in a sentence onto spatial calculi is the second step in our framework. We denote this as *spatial formalizing task*.
Task definition. We define spatial formalizing as *the automatic mapping of the output of SpRL to formal relations in spatial calculi*. In the previous section we have assumed that our spatial role representation covers all the spatial semantic aspects according to HSS. For the target representation of spatial formalizing we also require that it can express various kinds of spatial relations.
Spatial challenges. Ambiguity and under-specification of spatial information conveyed in the language, but also overspecification of spatial calculi models, make a direct mapping between the two sides difficult [2]. Most of the qualitative spatial models focus on a single aspect, e.g. topology, direction, or shape [12]. This is a drawback, particularly from a linguistic point of view and with respect to the pervasiveness of the language. Hence spatial formalizing should cover multiple aspects with practically acceptable level of generality. In the work of [5] the alignment between the linguistic and logical formalizations is discussed. Since these two aspects are rather different and provide descriptions of the environment from different viewpoints, constructing an intermediate, linguistically motivated ontology is proposed to establish a flexible connection between them.

GUM (*Generalized Upper Model*) is the state-of-the-art example of such an ontology [1, 17]. Moreover, in [5] $S-$connections are suggested as a similarity-based model to make a connection between various formal spatial systems and mapping GUM to various spatial calculi. However, obtaining an annotated corpus is the main challenge of machine learning for mapping to the target relations/ontology. In this respect using an intermediate level with a fairly large and fine-grained division of concepts is to some extent difficult and implies the need to have a huge labeled corpus. In addition, the semantic overlap between the included relations in the large ontologies makes the learning model more complex.

Moreover, mapping to spatial calculi is an inevitable step for spatial reasoning. Hence even if a corpus is constructed by annotating with a linguistically motivated ontology, mapping to spatial calculi still should be handled as a separate and difficult step. Even at this level, it is not feasible to define a deterministic mapping by formulating rules because bridging models to each other is not straightforward and external factors, context and all the involved spatial components, discourse features, etc influence this final mapping. Therefore the relationships between instances in different domains are not deterministic and they are often ambiguous and uncertain [5]. Given that for each learning step, a corpus should be available, we argue that it seems most efficient to learn a mapping from SpRL to (one or several) spatial calculi directly.

Representation based on spatial calculi. To deal with these challenges we proposed an annotation framework [10] inspired by the works of SpatialML [14] and a related scheme in [18]. We suggest to map the extracted spatial indicators and the related arguments onto the general type of the related spatial relation Region, Direction, Distance because these relations cover all coarse-grained aspects of space (except shape). The specific relation expressed by the indicators is stated in the suggested scheme with an attribute named *specific-type*. If the general-type is REGION then we map this onto topological relations in a qualitative spatial reasoning formalism, so the specific-type will be RCC8 which is a popular formal model. For directions the specific type gets a value in {ABSOLUTE, RELATIVE}. For absolute directions we use {S(south), W(west), N(north), E(east), NE(northeast), SE(southeast), NW(northwest), SW(southwest)} and for relative directions {LEFT, RIGHT, FRONT, BEHIND, ABOVE, BELOW} which can be used in qualitative direction calculi. Distances are tagged with {QUALITATIVE, QUANTITATIVE} (cf. [10]). To provide sufficient flexibility in expressing all possible spatial relations our idea is to allow more than one formal relation to be connected to one linguistic relation, helped by a (probabilistic) logical representation. The following examples illustrate this.

a)...and next to that left of that is my computer, perhaps a meter away.
Let `X`=my computer, `Y`=that, then a SpRL gives `nextTo(X, Y)`, `leftOf(X, Y)`, and a resulting spatial formalization is `DC(X, Y)`, `LEFT(X, Y)`, `Distance(X, Y,'value')` which in GUM corresponds to `leftprojectionexternal`.
b) The car is between two houses.
SpRL: `between(car, houses)`, spatial relations: `left(car, houses)` AND `right(car, houses)` which corresponds to GUM's `Distribution`.

c) The wheatfield is in line with crane bay.
SpRL: `inline(wheatfield, cranebay)`, spatial relations: `behind(wheatfield, cranebay)` XOR `front(wheatfield, cranebay)` GUM: `RelativeNonProjectionAxial`

Approach. The above mentioned examples show that a logical combination of basic relations can provide the required level of expressivity in the language. These annotations will enable learning probabilistic logical models relating linguistic spatial information to relations in multiple spatial calculi. Afterwards qualitative (or even probabilistic) spatial reasoning will be feasible over the produced output. The learned relations could be considered as probabilistic constraints about most probable locations of the entities in the text. Probabilistic logical learning [3] provides a tool in which considerable amounts of (structured) background knowledge can be used in the presence of uncertainty. The available linguistic background knowledge and features includes i) the features of the first step of spatial role labeling (syntactic, lexical and semantical information from the text) and ii) linguistic resources such as WordNet, FrameNet, language models and word co-occurences [20]. These could be combined with visual features extracted from visual resources in a multimodal environment for more specification of spatial relations. Structured outputs (i.e. the mapping to formal relations) could be learned in a joint manner. By exploiting a joint learning platform, annotating a corpus by aforementioned spatial semantics in addition to annotating by the final spatial relations (derived from spatial calculi) is less expensive than annotating and learning the two levels independently. Implementing such a learning setting is ongoing work.

4 Current Experiments

To start with empirical studies, we have performed experiments on the first SpRL learning phase. We learn to identify spatial indicators and their arguments trajector and landmark. We do not treat *motion*, *path* and *frame of reference* in this paper, and focus solely on prepositions as spatial indicators here.
Spatial preposition. For unrestricted language it seems valuable to first *recognize whether* there is any spatial indicator in the text. Since prepositions mostly play key roles for the spatial information in the first step we examine whether an existing preposition in a sentence conveys a spatial sense. Here we use linguistically motivated features, such as parse and dependency trees and semantic roles. We extracted these features from the training and test data of the TPP data set and tested several classifiers. The current results are a promising starting point for the spatial sense recognition and the extraction of spatial relations. The selected features were evaluated experimentally and our final coarse-grained MaxEntropy sense classifier outperformed the best system of the SemEval-2007 challenge by providing an F1 measure of about 0.874. We achieved an accuracy of about 0.88 for the task of recognizing whether a preposition has a spatial meaning in a given context.
Extraction of trajector and landmark. In the second SpRL step we extract the trajector and landmark arguments. Our features are inspired by those

in SRL. The main difference is that the pivot of the semantic relations here is the preposition, and not the predicate. The features from the parse/dependency tree and semantic role labeler are extracted from GUM examples. We labeled the nodes in the parse tree with GUM labels `trajector`(`locatum`), `landmark`(`relatum`) and spatial indicator (`spatialModality`).

We assume the spatial indicator (preposition) is correctly disambiguated and given, i.e. we perform a multi-class classification of parse tree nodes by `trajector`, `landmark` and `none`, for which we employed standard classifiers (naive Bayesian (NB), and maximum entropy (MaxEnt)). In addition, we tagged the sentences as sequences using the same features and applied a simple sequence tagger based on conditional random fields (CRF). The spatial annotations of GUM were altered in some instances to be able to obtain more regular patterns for machine learning. We labeled the continuous words (prepositions) and their modifiers as one spatial modality even if they had been tagged as individual relations in GUM, and we do not tag implicit trajectors/landmarks. In ongoing experiments we classify the headwords instead of whole constituents [9]. Table 1 presents the preliminary results for "trajector" (T) and "landmark" (L) recognition including overall accuracy evaluated by 10-fold cross validation. The simple multi-class classification ignores the global correlations between classes and as Table 1 indicates, more sophisticated CRF models can improve the results in particular for landmarks. Since the main sources of errors are a lack of data and the dependency of spatial semantics on lexical information, we will employ additional (lexical) features and ideally will use a larger corpus in our future experiments. However the current results show the first step of applying machine learning for SpRL and indicate a promising start towards achieving the entire automatic mapping from language to spatial calculi.

Method	F1(T)	F1(L)	Acc(All)
NBayes	0.86	0.70	0.94
MaxEnt	0.91	0.767	0.965
CRF	0.928	0.901	0.921

Table 1. Extraction of trajector (T) and landmark (L)

5 Conclusion and Future Directions

We have introduced a model for mapping natural language to spatial calculi. Both aspects of *spatial role labeling* and *spatial formalizing* have been described. A number of related problems that cause difficulties and ambiguities were addressed, and we have shown preliminary results for experiments on SpRL and the extraction of trajectors and landmarks. Our main idea for future work is to obtain (i.e. create) a corpus which is labeled by holistic spatial semantics plus a combination of spatial calculi. Each relation in the language can be connected to a *set* of relations belonging to predefined spatial calculi. This gives a logical representation of the language based on spatial calculi. We aim to learn statistical relational models for this. This enables adding probabilistic background knowledge related to structural information and spatial semantic notions, and supports (probabilistic) spatial reasoning over the learned models.

References

1. J. Bateman, T. Tenbrink, and S. Farrar. The role of conceptual and linguistic ontologies in discourse. *Discourse Processes*, 44(3):175–213, 2007.
2. J. A. Bateman. Language and space: a two-level semantic approach based on principles of ontological engineering. *Int. J. of Speech Tech.*, 13(1):29–48, 2010.
3. L. De Raedt, P. Frasconi, K. Kersting, and S. Muggleton, editors. *Probabilistic Inductive Logic Programming*, volume 4911 of *LNCS*. Springer, 2008.
4. Antony Galton. Spatial and temporal knowledge representation. *Journal of Earth Science Informatics*, 2(3):169–187, 2009.
5. J. Hois and O. Kutz. Counterparts in language and space – similarity and s-connection. In *Proceedings of the 2008 Conference on Formal Ontology in Information Systems*, 2008.
6. J. D. Kelleher. *A Perceptually Based Computational Framework for the Interpretation of Spatial Language*. PhD thesis, Dublin City University, 2003.
7. J. D. Kelleher and F. J. Costello. Applying computational models of spatial prepositions to visually situated dialog. *Comput. Linguist.*, 35(2):271–306, 2009.
8. T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward understanding natural language directions. In *HRI*, 2010.
9. P. Kordjamshidi, M van Otterlo, and M. F. Moens. Spatial role labeling: Automatic extraction of spatial relations from natural language. Technical report, Katholieke Universiteit Leuven, 2010.
10. P. Kordjamshidi, M van Otterlo, and M. F. Moens. Spatial role labeling: task definition and annotation scheme. In *LREC*, 2010.
11. H. Li, T. Zhao, S. Li, and J. Zhao. The extraction of trajectories from real texts based on linear classification. In *Proceedings of NODALIDA*, 2007.
12. W. Liu, S. Li, and J. Renz. Combining RCC-8 with qualitative direction calculi: algorithms and complexity. In *IJCAI*, 2009.
13. K. Lockwood, K. Forbus, D. T. Halstead, and J. Usher. Automatic categorization of spatial prepositions. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006.
14. I. Mani, J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner. SpatialML: Annotation scheme, corpora, and tools. In *LREC*, 2008.
15. L. Màrquez, X. Carreras, K. C. Litkowski, and S. Stevenson. Semantic role labeling: An introduction to the special issue. *Comp. Ling.*, 34(2):145–159, 2008.
16. R. J. Mooney. Learning to connect language and perception. In *AAAI*, 2008.
17. R. Ross, H. Shi, T. Vierhuff, B. Krieg-Brückner, and J. Bateman. Towards dialogue based shared control of navigating robots. In *Proceedings of Spatial Cognition IV: Reasoning, Action, Interaction*, pages 478–499, 2005.
18. Q. Shen, X. Zhang, and W. Jiang. Annotation of spatial relations in natural language. In *Proceedings of the International Conference on Environmental Science and Information Application Technology*, 2009.
19. D. A. Tappan. *Knowledge-Based Spatial Reasoning for Automated Scene Generation from Text Descriptions*. PhD thesis, New Mexico State University Las Cruces, New Mexico, 2004.
20. M. Tenorth, D. Nyga, and M. Beetz. Understanding and Executing Instructions for Everyday Manipulation Tasks from the World Wide Web. In *ICRA*, 2010.
21. J. Zlatev. Spatial semantics. *In Hubert Cuyckens and Dirk Geeraerts (eds.) The Oxford Handbook of Cognitive Linguistics, Chapter 13*, pages 318–350, 2007.

Human evaluation of robot-generated spatial descriptions

Simon Dobnik and Stephen G Pulman

Computing Laboratory, University of Oxford
Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom
{simon.dobnik,stephen.pulman}@comlab.ox.ac.uk
http://www.clg.ox.ac.uk

Abstract. We describe a system where the semantics of spatial referential expressions have been automatically learned by finding mappings between symbolic natural language descriptions of the environment and non-symbolic representations from the sensory data of a mobile robot used for localisation and map building (SLAM). Although the success of learning can be measured by examining classifier performance on held-out data, this does not in itself guarantee that the descriptions generated will be natural and informative for a human observer. In this paper we describe the results of an evaluation of our embodied robotic system by human observers.

Key words: spatial expressions, machine learning, mobile robots, embodied multi-modal conversational agents, evaluation

1 Introduction

A conversational robot must be able to refer to and resolve references to the environment in which it is located with its human conversational partner. Mapping between the linguistic and non-linguistic representations is commonly performed by first identifying some parameters of the physical world on the basis of psychological evidence and then integrating them into customised functions [1, 2]. However, in a real robotic system which has been primarily built for tasks such as map building, localisation and navigation the information required by such models may not be readily available. Our approach attempts to use a simple model of space and motion that is available to a mobile robot and show that a mapping between its representations and highly abstract natural language spatial descriptions can be learned: that the robot can display a human-like performance in "understanding" and generating spatial descriptions or motion in new environments. In this paper we focus on the evaluation of the robot's performance from the point of view of a human conversational partner.

2 Learning spatial descriptions

Spatial descriptions may be about the identity of objects in a scene [3], about the spatial relations between the objects in a scene [4] or about the route that a moving

object can take in a scene [5]. The scene may be a small artificial town on a table top, a building with rooms or a real town. Our scenario is a larger room, a lab, which is constrained by walls, and which contains life-sized objects such as a chest, a box, a table, a pillar, a stack of tyres, a chair, a desk and shelves. The natural language descriptions that can be made in this environment belong to two categories: they can be descriptions of the robot's motion such as "You're going forward slowly" or, when the robot is stationary, descriptions of relations between the objects in the scene "The table is to the left of the chair". We consider descriptions of motion spatial description because their meaning is also relative to the environment in which they are used.

We use an ATRV-JR mobile robot designed by iRobot which runs middle-ware called MOOS.[1] The system runs an odometry component which provides information about the robot's motion such as its ⟨R-Heading⟩[2] and ⟨Speed⟩ and the SLAM localisation component [6] which uses a previously built 2-dimensional SLAM map to localise the robot. The objects were grounded on the map manually by taking a centre point of the cloud of points representing them, for example: chair ⟨0.6234,0.2132⟩ (⟨X⟩ and ⟨Y⟩). Our representation of the state of the robot and the space around it is thus extremely simple but the values of such representations are very accurate.

A group of four non-expert volunteers was invited to provide linguistic spatial descriptions of the robot and its environment. Each was first familiarised with the scene, the names of the objects and the different types of motion that the robot can produce. Then they were instructed to describe the motion and the location of objects from the perspective of the robot. This ensured that all directionals were used unambiguously from a single reference frame [7]. Two datasets were created. The linguistic descriptions in the first dataset (*Simple*) were made by a single describer and were restricted to a pre-defined small vocabulary (16 words) that appeared as choices on a computer screen. The second dataset (*All*) was created by all four participants who could use unrestricted vocabulary and sentences. Such descriptions show considerable lexical variation (46 words) but their syntactic structure is limited and in most cases similar to the examples above.[3] The two settings were intended to show the effects of subjectivity on the datasets and the models produced. To preserve the naturalness of the situation we used speech recognition (with some consequent noise in the language).

To turn MOOS log files (where both linguistic and non-linguistic information was recorded) to learning instances a few processing steps had to be performed: the locations of objects were expressed relative to the robot (rather than being global values relative to some random point where the robot has started) and their values were normalised (given the estimated size of the room or the maximum speed of the robot in the current session). This ensured that the models that were built could be later applied to new contexts. Words from natural language descriptions

[1] MOOS was designed by Paul M. Newman (Mobile Robotics Group, Department of Engineering, University of Oxford). We would like to thank him and members of his group for introducing us to mobile robotics.

[2] The attributes used in learning are marked with angled brackets.

[3] Complex descriptions such as "the chair is to the left of the table and behind the sofa" were simplified as two descriptions of relation.

were tagged to one of the categories ⟨Verb⟩, ⟨Direction⟩, ⟨Heading⟩, ⟨Manner⟩ and ⟨Relation⟩ which were also the target classes to be learned. The learning was accomplished with the Weka toolkit [8] which includes a range of offline supervised classifier implementations and a common framework to represent the data and evaluate the results. Each of the target linguistic classes was learned separately and not all attributes were used in each learning exercise. For example, to learn the category of ⟨Verb⟩ we only used the ⟨R-Heading⟩ and ⟨Speed⟩ attributes and to learn the category ⟨Relation⟩ we used the attributes ⟨LO_x⟩, ⟨LO_y⟩, ⟨REFO_x⟩ and ⟨REFO_y⟩ where LO stands for a located object and REFO stands for a reference object. Including all attributes resulted in a considerably lower classifier accuracy since may spurious relations were discovered.

The classifiers that were used in the human evaluation experiments described in the following sections were produced by the J48 learner which is the Weka's implementation of the ID3/C4.5 decision tree learner [9]. Their estimated accuracies obtained by a stratified 10-fold cross-validation are given in the last column of Table 1 for both *Simple* and *All* datasets. Note that these values are not the best values that we obtained. The accuracy of the motion categories was improved by a better method of combining a set of temporally sequential observations from the robotic log to instances. We also compared the performance of different machine learning methods on our datasets.

3 Evaluation by humans

The evaluation of machine learning classifiers by a stratified 10-fold cross-validation tests the degree to which the descriptions learned will generalise correctly to new cases. However, it does not tell us whether the models that are built will result in linguistic behaviour natural to humans. In order to know this we carried out a user study. We integrated the classifiers to a simple system that generates descriptions called *pDescriber*. This considers the current (normalised) values of the same attributes that were used in learning and predicts linguistic target classes. If the robot is moving, it generates descriptions of motion; if it is stationary it generates descriptions of object relations. The values of the predicted categories are applied to syntactic patterns such as "I'm ⟨Verb⟩ing" or "⟨LO⟩ is ⟨Relation⟩ the ⟨REFO⟩" which produce sentences that are subsequently pronounced by a speech synthesiser, for example "I'm reversing" and "You are behind the chair".

A new room was set up. Most of the objects were the same as in the data collection exercise but their placement was different. Five subjects were invited to the lab for approximately an hour each. None of them had participated in data collection. After being introduced the scene, they were explained that they should indicate whether they agree with the description that was generated by the robot given its current state and that of the environment. This gave us simple binary data. If they disagreed with the description, they had a chance to provide a better description. Note that the descriptions were not evaluated as utterances but per linguistic category. For example, for each utterance the system would query the evaluator whether "right" was a good word to describe the robot's heading in which it was moving or whether "to the left of" was a good description of the relation

between the chair and the table. The evaluators were also invited to make qualitative judgements about the appropriateness of the descriptions which we noted down. For approximately one half of the session the system used the classifiers built from the *Simple* dataset and the other half it used the classifiers built from the *All* dataset.

4 Evaluator-system agreement

The central part of Table 1 shows the measured accuracies from each evaluator per category. As explained in the previous section, accuracy is measured as evaluator agreement with the system on the choice of description. The penultimate column contains the accuracies when all evaluators are considered together. The last column contains the estimated accuracies of the classifier that the system was using to produce these descriptions. The table is split into two parts each containing the results from one configuration of the system (*J48-Simple* and *J48-All*).

Table 1. System performance *vs.* classifier performance

Category		Evaluators						Classifier
		a	b	c	d	e	All	J48
								Simple
Motion	$n =$	36	17	14	2	21	90	–
	Verb	100	**88.24**	100	100	95.24	96.67	89.02
	Direction	100	**76.47**	100	100	100	95.56	87.80
	Heading	100	**82.35**	100	100	**85.71**	**93.33**	97.56
	Manner	100	82.35	100	100	100	96.67	70.73
Relation	$n =$	65	23	19	53	22	182	–
	Relation	**67.69**	**65.22**	**68.42**	**66.04**	**59.09**	**65.93**	75.90
								All
Motion	$n =$	53	22	53	7	41	176	–
	Verb	96.23	77.27	88.68	100	100	92.61	48.22
	Direction	96.23	72.73	92.45	100	100	93.18	55.68
	Heading	98.11	68.18	92.45	100	95.12	92.05	60.77
	Manner	100	72.73	98.11	100	100	96.02	54.70
Relation	$n =$	66	28	72	110	58	334	–
	Relation	72.73	**57.14**	**44.44**	70.00	**43.10**	**59.28**	69.12

How do the results from the evaluation of the system by humans and the evaluation of the underlying classifiers compare? The classifier accuracies are the average accuracies obtained through a 10-fold cross-validation. In the human evaluation of the system the accuracy is determined on an independent test set. In both cases the reported accuracy is the ratio between the number of agreements with the system or correct classifications over the total number of considered testing instances. There is a slight difference between the two situations in how a positive match is made. In cross-validation the correct value of the class is pre-defined and hidden from the classifier and this is matched with the predicted class. In human evaluation an evaluator hears the generated description before they give their evaluation.

This description is the one that is predicted by a classifier given the attributes representing the robot's current internal state. In this respect it is possible that the system unavoidably biases the evaluator, since other possible descriptions are never produced. Furthermore, when evaluating the system in this way, the observers are not always just evaluating the classifiers. For example, when generating descriptions of object relations the located and the reference objects are chosen at random and the classifier is used to predict the best relation between the two. The description may be evaluated as unsuitable because of an unfortunate choice of objects even though the spatial relation between them is correct.

A quick look at the table reveals that the evaluators considered the system performance to be better than the accuracy of the underlying classifiers on most classes of motion descriptions (J48-Simple classifier: $\bar{x} = 86.28\%$, J48-Simple evaluators: $\bar{x} = 95.56\%$; J48-All classifier $\bar{x} = 54.84\%$, J48-All evaluators: 93.47%). To make the comparison easier we mark the values where the *opposite* is true, when the system is evaluated to perform worse than its classifiers, in bold. The evaluator accuracies are quite similar across categories, even for the ⟨Manner⟩ category on which the the classifiers perform less well than others. This is more the case with the *Simple* configuration than *All*. On the contrary, the system was considered to perform less well than its classifiers on the ⟨Relation⟩ category by approximately 10% in both cases (J48-Simple classifier: 75.90%, J48-Simple evaluators: 65.93%; J48-All classifier: 69.12%, J48-All evaluators: 59.28%).

The scores from evaluator b are lower than those from other evaluators, particularly on the motion classes and for the *J48-Simple* configuration. The numbers in lines starting with $n =$ indicate the size of the evaluation sample. Although the number of descriptions that the robot generated was not strictly controlled, a reasonable sample was obtained for each evaluator. The only exception is evaluator d who evaluated only a small number of descriptions of motion but on the other hand considered more descriptions of object relations.

An explanation why the evaluators consider the system to perform better than its underlying classifiers on the motion categories but not on the relation category could be that motion categories contain words that are less semantically restrictive. For example, the category ⟨Verb⟩ contains words such as "going", "moving" and "continuing" which all have a very similar reference for a human but not for a machine learner where the attribute values are assumed to be discrete. Consequently, an evaluator may accept such alternative. The categories ⟨Direction⟩, ⟨Heading⟩ and ⟨Manner⟩ contain words with clearer semantic divisions but they all also contain a word "none" which was assigned as a value of each category in machine learning dataset if a word for that category was not present. The meaning of this word is ambiguous between a default meaning and an anaphoric meaning. For the ⟨Direction⟩ category "none" has the same meaning as "straight". However, it can also refer anaphorically to the previously generated description of direction if this has not changed.

Another explanation why the results are different for descriptions of motion and object relations is that learning and generating of the latter is more complex. It could be that our learning and generation models for descriptions of object relations capture human knowledge less well than the models for description of motion. We discuss some qualitative evidence for this in Section 6.

5 Inter-evaluator agreement

Agreement between individual evaluators demonstrates that the system has not been tuned to the vocabulary of the describers who provided descriptions for machine learning. Disagreement may be informative too: if evaluators collectively disagree it means that the generation task is not subjective, that there exists a consensus on what is a good description in a particular context and what is not.

Unfortunately, inter-evaluator agreement cannot be established directly, for example by calculating a κ coefficient, because not all evaluators evaluated the same set of items. The evaluators considered a closed set of words produced by the system. We can expect that the agreement of a single evaluator with the system will not be identical on every word that it produces. Some words are more difficult to learn than others. If so, the difference in the ratings for words should be consistent across evaluators. According to our model of agreement, an evaluator agrees with other evaluators if their accuracy scores per word correlate with the mean of accuracy scores per word of everyone else.

Table 2. Agreement of each evaluator with the rest of the group

Configuration	a:rest	b:rest	c:rest	d:rest	e:rest	Mean
J48-Simple	0.824**	0.382 ns	0.787**	0.907**	0.636*	0.707
J48-All	0.504*	0.048 ns	0.635**	0.756**	0.662**	0.521

Table 2 shows the Pearson's correlation coefficients r_{xy} obtained at each fold of correlation for both sets of classifiers. The last column contains the average correlation coefficient. The asterisks indicate the statistical significance levels of the coefficients obtained by a two-tailed t-test.[4]

We can see that except for the evaluator b there exists a moderate to high correlation between the scores of an individual evaluator and the mean scores of the rest of the group. The average correlation coefficient for the *J48-Simple* configuration is greater (0.707) than the average correlation coefficient for the *J48-All* configuration (0.521). All correlation coefficients, except in the case of evaluator b are statistically significant at the level $\alpha = 0.05$. In sum, apart from evaluator b there is a considerable consensus between the remaining four evaluators on the performance of the system. Thus, it has captured some universal knowledge.

6 Qualitative evaluation

Descriptive observations made by the evaluators are useful because they point out facts about spatial cognition and the shortcomings of the system that can be further improved [10, 11].

[4] * indicates that the correlation is significant at the 0.05 level, and ** indicate that it is significant at the 0.01 level. "ns" indicates that the correlation is not significant.

Ambiguity of heading and direction. The descriptions such as "left" and "right" are ambiguous when used to refer to motion. "Moving right" can mean moving forward with a heading in the clockwise direction. It can also mean making a sudden turn to the region that is to the right of the current location and then moving straight in that direction. Similarly, "moving backward" can mean that the robot is moving in the direction that is behind its back (reversing) or that it has reversed but is now moving forward in the direction that was previously behind its back. The second of each description pair is more complex and to learn such descriptions the learner would have to abstract over a set of actions rather than over physical descriptions of environment. Since while performing the second action "right" and "backward" may refer to the same state of the robot as "straight" and "forward" in our model, the robot is likely to over-generate such descriptions in cases where the first action was not performed.

Object shape. The SLAM map used in our model does not contain abstract representations of objects but only clouds of points. Each object is represented by a centre point. While this works reasonably well for objects that square-shaped, difficulties arise with objects that are markedly different in one dimensions such as "the wall" and "the barrier".

Switching the reference frame. Although evaluators were told that the descriptions generated with the reference frame fixed on the robot or from "its perspective", it was very easy for them to switch from this relative reference frame to the intrinsic reference frame fixed on the reference object. Firstly, it became apparent that some switches to the intrinsic reference frame have been learned from the training data and such descriptions appeared appropriate in the current context. In this case, the majority of evaluators would accept such descriptions although they should not do so according to our instructions. Secondly, properties of some objects invite human describers or observers to use intrinsic rather than the relative reference frame. This is true for objects that are larger than describer (walls, barriers and cupboards), have an identifiable front and are animate (another robot). Only the intrinsic reference frame is possible when the robot describes its own location and consequently cannot serve as a reference object. "I'm in front of the chair" unambiguously means that the robot is located in the region around and orientated by the seating area of the chair. Note that the reference frame also applies to the projective descriptions of motion.

Reference to objects outside the robot's field of vision. There was a disagreement between the evaluators whether descriptions that cannot be "seen" by the robot are appropriate or not. Technically, "the vision field" of the robot is much greater than that of a human observer - it is the entire SLAM map which represents its mental map. Humans also use mental maps to imagine configurations of objects for tasks such as navigation and therefore descriptions of objects not in the visual focus of the describer may not be completely unnatural. In fact, particularly disapproved were those descriptions where only one of the objects was "visible".

Non-optimal choice of objects. The classifiers always attempt to predict the best description of relation between two objects and may do so but the description may

be judged inappropriate because of an unfortunate selection of objects. The latter can be accomplished by a contextual model which our system does not implement. Given that we are primarily interested in spatial relations itself the choice of objects at random seems to be reasonable. Some evaluators were more sympathetic to such descriptions than others. However, they all agreed that descriptions where the lack of object salience was coupled with the lack of the vision field salience were quite unacceptable.

7 Conclusion

Although our classifiers use a relatively simple (topological) representation of space primarily intended for localisation of a mobile robot we can conclude that they work surprisingly well in practice in replicating human linguistic competence. They fall short sometimes because they do not have access to non-topological information such as object shape, reference frame, discourse structure for modelling salience and world knowledge about the objects. Such data must be provided from other sources.

References

1. Regier, T., Carlson, L.A.: Grounding spatial language in perception: an empirical and computational investigation. Journal of Experimental Psychology: General 130(2), 273–298 (2001)
2. Coventry, K.R., Cangelosi, A., et al.: Spatial prepositions and vague quantifiers: implementing the functional geometric framework. In: Freksa, C., Knauff, M., et al. (eds.) Spatial Cognition, vol. IV, pp. 98–110. (2005)
3. Zender, H., Martínez-Mozos, O., et al.: Conceptual spatial representations for indoor mobile robots. Robotics and Autonomous Systems 56(6), 493–502 (2008)
4. Roy, D.K.: Learning visually-grounded words and syntax for a scene description task. Computer speech and language 16(3), 353–385 (2002)
5. Lauria, S., Kyriacou, T., et al.: Converting natural language route instructions into robot-executable procedures. In: Proceedings of Roman'02. pp. 223–228. (2002)
6. Bosse, M., Zlot, R.: Map matching and data association for large-scale two-dimensional laser scan-based SLAM. IJRR 27(6), 667–691 (2008)
7. Steels, L., Loetzsch, M.: Perspective alignment and spatial language. In: Coventry, K.R., Tenbrink, T., Bateman, J. (eds.) Spatial language and dialogue, Explorations in Language and Space, vol. 3, pp. 70–88. OUP (2009)
8. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2nd edn. (2005)
9. Quinlan, J.: C4.5: programs for machine learning. Morgan Kaufmann (1993)
10. MacMahon, M., Stankiewicz, B., Kuipers, B.: Walk the talk: Connecting language, knowledge, and action in route instructions. In: Proceedings of AAAI-2006. pp. 1475–1482. (2006)
11. Moratz, R., Tenbrink, T.: Spatial reference in linguistic human-robot interaction: iterative, empirically supported development of a model of projective relations. Spatial Cognition & Computation 6(1), 63–107 (2009)

Connecting Language and Geography with Region-Topic Models

Michael Speriosu, Travis Brown, Taesun Moon, Jason Baldridge, and Katrin Erk

The University of Texas at Austin
Austin TX 78712, USA
{speriosu,travis.brown,tsmoon,jbaldrid,katrin.erk}@mail.utexas.edu

Abstract. We describe an approach for connecting language and geography that anchors natural language expressions to specific regions of the Earth, implemented in our *TextGrounder* system. The core of the system is a region-topic model, which we use to learn word distributions for each region discussed in a given corpus. This model performs toponym resolution as a by-product, and additionally enables us to characterize a geographic distribution for corpora, individual texts, or even individual words. We discuss geobrowsing applications made possible by TextGrounder, future directions for using geographical characterizations of words in vector-space models of word meaning, and extending our model to analyzing compositional spatial expressions.

Keywords: Geobrowsing, Toponym Resolution, Topic Models

1 Introduction

Incredible amounts of text are now readily available in digitized form in various collections spanning many languages, domains, topics, and time periods. These collections are rich sources of information, much of which remains hidden in the sheer quantity of words and the connections between different texts. Techniques that reveal this latent information can transform the way users interact with these archives by allowing them to more easily find points of interest or previously unnoticed patterns. In this paper, we describe our preliminary progress in developing our *TextGrounder* system, which we use to create geospatial characterizations and visualizations of text collections. We also discuss the potential for using the representations produced by our system to inform or learn models of how language encodes spatial relationships.

The spatial meaning of an utterance depends on many factors. The expression *a barbecue restaurant 60 miles east of Austin* has a compositional analysis in which one must: (1) identify whether *Austin* refers to a person or place and which person or place it is, including determining the correct latitude and longitude associated with it; (2) identify the location that is 60 miles to the east of that location; and (3) possibly identify a restaurant that serves barbecue in that vicinity. We do not tackle such compositional analysis yet; instead we begin with a standard bag-of-words model of texts that allows us to use the geographic focus

of words like *barbecue* and *restaurant* and other terms in the document to disambiguate (potential) toponyms like *Austin* and landmarks like *the Eiffel Tower*.[1] Our model *learns* that locations are highly associated with certain vocabulary items without using labeled training material; it relies only on a gazetteer. To do this, we use a simple topic model [2] that construes regions of the Earth's surface as topics. We refer to this as the *region-topic model*.

There are at least two linguistically interesting outcomes that could arise from this modeling strategy. The first is that it directly provides a light-weight form of grounding natural language expressions by anchoring them to (distributions over) locations on the Earth. This presents an opportunity to add spatially relevant features into recent vector space models of word meaning (e.g. [4]). Typically, the dimensions of vector space models are not interpretable, and the only way that a vector representation of a word can be interpreted is through its distance to the vectors of other words. In contrast, dimensions relating to locations on Earth will be informative and interpretable in themselves. This will allow us to explore the question of whether such vector space models support additional inferences informed by world knowledge. Second, our approach is language independent, and the fact that expressions are grounded geographically presents the opportunity—without using labeled data, e.g. as with SpatialML [9]—to eventually learn the meaning of expressions like *X 60 miles east of Y*, based on texts that express many different referential noun phrases *X* and *Y*, some of which will be locations which we can resolve accurately.

We aim to use TextGrounder to improve information access for digitized text collections. We are working with a collection of ninety-four British and American travel texts from the nineteenth and early twentieth centuries that were digitized by the University of Texas libraries.[2] These texts are replete with references to locations all around the Earth, so they are an ideal target for geobrowsing applications (e.g. in Google Earth) that display the relative importance of different locations and the text passages that describe them. This kind of analysis could be used to provide "distant reading" interfaces for literary scholarship [12], to support digital archeology [1], or to automatically produce geographic visualizations of important historical events, such as mapping survivor testimonies of the Rwandan genocide. It could also enable users to create mashups of temporally and generically diverse collections, such as Wikipedia articles about the Civil War with contemporary accounts by soldiers and narratives of former slaves.

2 System

TextGrounder performs *geolocation* in a very general sense: it connects natural language texts, expressions, and individual words to geographical coordinates and distributions over geographical coordinates. The most basic and concrete application of geolocation is *toponym resolution*, the identification and disambiguation of place names [7]. For instance, there are at least forty places around

[1] Which could be in Paris (France), Paris (Texas), Las Vegas (Nevada), etc.

[2] `http://www.lib.utexas.edu/books/travel/index.html`

the world called *London*; a toponym resolver must identify that a particular mention of London refers to a place (and not a person, like *Jack London*) and identify which *London* was intended as the referent (e.g., London in Ontario or England). Most systems focus solely on recognizing the places associated with texts based on matching known names to known locations. Typically, simple pattern matching or heuristics are used to identify and disambiguate places.

TextGrounder performs toponym resolution as a by-product; it automatically interprets references to places, landmarks, and geographic features in free text, and uses that information to provide location information on digital maps. Because it learns from raw text, the system uses information and representations that support a much more general connection between language and geography than toponym resolution alone. The system thus performs a light-weight form of grounding computational representations of words in the real world.

The underlying model, depicted in Figure 1, is an adaptation of probabilistic topic models [2]. Topics are simple distributions over the vocabulary for which some particular words have higher probability than others—for example, a topic related to sports would have high probability for words like *team*, *game*, and *ball*. To adapt this approach for geolocation, we represent the Earth as a set of non-overlapping 3-by-3 degree regions, where each region corresponds to a topic. Each document is thus a mixture of region-topics, so different locations discussed in the same document can be modeled. Ultimately, this means that we associate word distributions with specific locations such that words that are more relevant to that location have higher probability. We do not retain all region-topics; instead, given a gazetteer, such as World Gazetteer[3], we consider only region-topics that spatially contain at least one entry in the gazetteer.

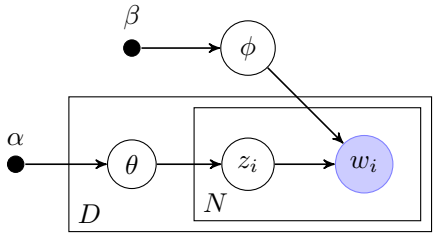

Fig. 1: Graphical representation of the region-topic model with plate notation. The N word observations w_i over D documents is conditioned on the word-level region assignments z_i and a word-by-region prior $\phi|z, \beta \sim \text{Dirichlet}(\beta)$. The topics are drawn from a multinomial on the region-by-document prior $\theta|d, \alpha \sim \text{Dirichlet}(\alpha)$ where $d \in D$. Structurally, the model is identical to a standard topic model—however, the initialization and interpretation of the topics is anchored by actual regions on Earth rather than arbitrarily assigned latent semantic concepts.

To analyze a corpus, we first run the Stanford named entity recognizer[4] (NER) and extract all expressions identified as locations. We then learn the region-topics for each word and toponym. Unlike standard topic models, where topics are not explicitly linked to an external representation, region-topics are

[3] `http://world-gazetteer.com/`
[4] `http://nlp.stanford.edu/ner`

anchored to specific areas of the Earth's surface. This allows us to initialize the inference procedure for our model by seeding the possible topics to only those for which we have some evidence; this evidence comes via toponyms identified by the NER system and the regions which contain a location indexed by those toponyms. The word distributions for non-toponyms in a text conditioned over regions are then inferred along with distributions for the region-constrained toponyms through a collapsed Gibbs sampler. Note that we do not consider the topology of the regions themselves (i.e. our model has no knowledge of the systems of neighborhoods which are inherent in the definition of regions over the globe); the present model is an intermediate step towards that goal.

Toponym resolution is performed implicitly by this model because the identified toponyms in a text are constrained to have positive joint probability only with the regions that enclose the corresponding, possibly ambiguous, coordinates in the gazetteer for those toponyms. If each toponym in a document is associated with multiple regions, the topic model will learn a topic and word distribution that assigns high probabilities to regions that coincide among the possible regions. For example, *London*, *Piccadilly* and *Hyde Park* might occur in the same document; each of these toponyms are ambiguously mapped to more than one region. There are different mixtures of regions that contain all these toponyms; the topic model will assign higher probability to an analysis that accounts for all of them in a single region (namely, the one containing London, UK). After a burn-in period for the Gibbs sampler, we take a single sample (or average over multiple samples) and geolocate the toponyms by placing the toponym on the coordinates which are resolved by the gazetteer and the region assignment.

The region-topic distributions include both toponyms and standard vocabulary items (non-toponyms). Because non-toponyms are unconstrained over regions, they provide additional evidence for determining the set of region-topics required to explain each document. Thus, they aid in toponym resolution *and* the model discovers the words that are most associated with each region. For example, the region-topic containing Austin, Texas would have high probability for words like *music*, *barbecue*, and *computers*, whereas for San Francisco, we'd expect *bay*, *finance*, and *tourism* to be prominent words. Based on these distributions, we can determine additional relationships, such as the distribution of a word over the Earth's surface (by considering its probability in each of the region-topics) or the similarity of different regions based on their corresponding region-topics (e.g. through information divergence measures).

3 Datasets and output

We seek to make the British and American travel collection more useful for scholars of the period through TextGrounder-generated KML (Keyhole Markup Language) files that may be loaded into a geobrowser like Google Earth, including (1) plotting the prominence of different locations on Earth in the collection, (2) embedding text passages at their identified locations for discovery, and (3) plotting the region-topic word distributions (see Figure 2). These preliminary

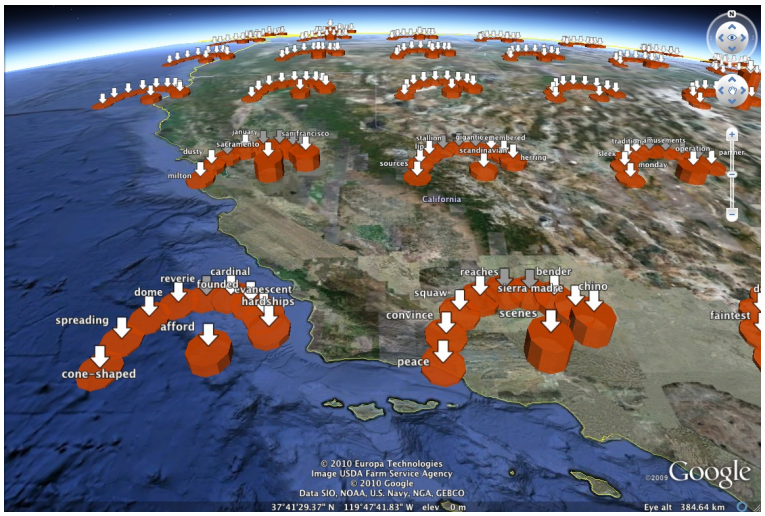

Fig. 2: TextGrounder visualization in Google Earth for John Beadle's *Western Wilds, and the Men who Redeem Them*. The top ten words associated with each region are shown, with red 3D bars indicating their relative frequency for that region.

visualizations provide subjective characterizations of the quality of the system output, which has been useful as we develop and refine our approaches. To obtain an objective measure of performance on the specific task of toponym resolution, we are now interfacing TextGrounder with the TR-CoNLL corpus [7], which contains 946 English-language newspaper articles that contain human-annotated ground truth for 5,963 toponym mentions.

To test the cross-lingual applicability of TextGrounder, we will create a multilingual geotagged subset of Wikipedia (see [13] for an extensive discussion of Wikipedia's geotagged articles and modeling based on them) that we can use as a test corpus. TextGrounder associates multiple regions with each document, but some regions will tend to dominate each document; we can thus choose a location that is most central to each document and check the geospatial distance from that location to the one annotated in Wikipedia. We will create the corpus by extracting pages in English, German, and Portuguese that have similar geographic coverage in each language (this is necessary because the English Wikipedia is much larger than the others and has more complete geotagging). We will identify a subset of pages in all three languages that discuss the same locations, using their geotagged information. This will be a reasonably large set: there are currently over 170,000 articles in English (and 1.2 million across all languages) that are annotated with a geotag for the main subject of the article.

The approach and methodology we advocate are general and flexible—the same methods can be applied relatively independently of the particular corpus being analyzed and the task at hand. The resulting robustness gives us confidence

that our approach will scale well, allowing us to provide geographical searching and browsing for a much wider range of documents than has been possible in traditionally curated literary or historical collections. The unsupervised methods we use allow a more useful mapping of texts because they do not base grounding entirely on toponyms; this means we can characterize the relative importance of different locations using a much wider array of evidence than those that simply resolve toponyms. Furthermore, incorporation of more diverse evidence is of retroactive benefit to toponym resolution, and we believe it will be mutually beneficial to jointly learn a textual hidden space and a geospatial model.

4 Spatial features and word meaning

Vector space models are a popular framework for the representation of word meaning, encoding the meaning of lemmas as high-dimensional vectors [6, 8]. In the default case, the components of these vectors measure the co-occurrence of the lemma with context features over a large corpus. Vector spaces are attractive because they can be constructed automatically from large corpora; however, the interpretation of the representation for a word is based solely on its distance in space to other words. The region-topic model provides an opportunity to represent the meaning of words through *grounded* features: words can be represented as a vector whose dimensions are region topics, and the coordinates are the word probabilities under the topics. This model overcomes the dichotomy of corpus-derived but uninterpretable versus human-generated and interpretable features: it is automatically derived, but offers directly interpretable geographical features.

We will use the region-topic models as a vector space model to study three sets of issues. (1) Traditional vector space models characterize the meaning of a word intra-textually, solely through other words. How do grounded representations compare on traditional tasks like word similarity estimation? Are they perhaps less noisy simply by virtue of pointing to extra-textual entities? (2) Similarity measures typically used in vector space models, such as Cosine and Jaccard, treat dimensions as opaque. In a model where dimensions are regions, we can exploit world knowledge in measuring similarity, for example by taking the distance between regions into account. Can this fact be used to derive better estimates of word similarity? (3) While most vector space models derive one vector per word, conflating senses of polysemous words, it is also possible to derive vectors for a word in a particular context [11, 3]. In a context of *eat apple*, the vector of *apple* would focus on the fruit sense of apple, suppressing features that speak to the company sense. This raises the question of whether it is possible to determine contextually appropriate interpretable features. In the example above, features like *Michigan*, *California* or *New Zealand* should be strengthened, while *Cupertino* (associated with Apple Inc.) should be suppressed. On the technical side, the main challenge will lie in the difference in strength between dimensions, due to different corpus frequencies of different senses of a polysemous word.

5 Related work

There has been quite a bit of research addressing the specific problem of toponym resolution (see [7] for an overview). Of particular relevance is the Perseus Project, which uses a heuristic system for resolving toponyms and creating automatically generated maps of texts written around the time of the Civil War [14].

The two current approaches that are most similar to ours are the location-aware topic model [10] and the location topic model [5], but the form of our model is different from both of these. The location-aware topic model assumes that every document is associated with a small set of locations, so its representation of geography is discrete and quite restricted. The location topic model is more similar to ours: they also seek to learn connections between words and geography using a topic model, and the visualizations they produce (for travel blogs) have a similar flavor. Interestingly, they model documents as mixtures of location-based topics and more general topics: this of course allows them to characterize words that do not have compelling specific geographical meaning. They preprocess their data, and perform toponym disambiguation using a heuristic system (the details of which are not given). Our model uses a different representation that actually grounds topics explicitly, because each topic is directly tied to a specific region on Earth. As a result, our model connects language to geography and performs toponym disambiguation as a by-product. We are interested in combining these two models to see how the learned word distributions differ and the effects they have on toponym disambiguation and our visualizations.

6 Conclusion

The Internet has become a repository of information in many of the world's languages, but the sheer quantity of written material—especially when considering multilingual contexts—also makes it harder to find or digest information of interest. We seek to create meaningful abstractions of language that allow large text collections to be browsed with respect to the places they discuss. These abstractions are learnable from unannotated texts, which greatly facilitates their use for any language with digitized material.

The historically and politically relevant collections that we are examining provide diverse materials that are replete with references to real people and places. This makes them an ideal target for geospatial resolution. Our model performs this resolution, but more importantly, it uses representations that enable many alternative ways of relating language to geography. This in turn supports many different ways to visualize texts geospatially, including seeing the geographic centrality of an entire collection or for a single word or expression, as well as exploring the text passages most relevant for a given location in context. These kinds of visualization will enable scholars to interact with massive text collections in novel ways, and will test the potential of maps to serve "not as all-encompassing solutions, but as generators of ideas" [12].

Additionally, these representations create the possibility to anchor natural language expressions to the real world in a light-weight fashion—this has the

potential to make them useful for inclusion in vector space models of word meaning. By starting at this level, using very simple assumptions about the dependencies between words (by treating texts as bags-of-words), we can analyze many texts and many languages. However, we ultimately are interested in deriving the geospatial meaning of *compositional* expressions—a very difficult task, but one which we hope our current models will help us eventually address.

TextGrounder is an ongoing effort. The system, example output and updated documentation are available on the project's website.[5]

Acknowledgments. We acknowledge the support of a grant from the Morris Memorial Trust Fund of the New York Community Trust.

References

1. Barker, E., Bouzarovski, S., Pelling, C., Isaksen, L.: Mapping an ancient historian in a digital age: the herodotus encoded space-text-image archive (hestia). Leeds International Classical Studies 9(1) (2010)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Erk, K., Padó, S.: A structured vector space model for word meaning in context. In: Proceedings of EMNLP. Honolulu, HI (2008)
4. Erk, K.: Representing words as regions in vector space. In: Proceedings of CoNLL-2009. pp. 57–65. Association for Computational Linguistics, Boulder, Colorado (June 2009)
5. Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.M., Pang, Y., Zhang, L.: Equip tourists with knowledge mined from travelogues. In: Proceedings of WWW 2010. pp. 401–410 (2010)
6. Landauer, T., Dumais, S.: A solution to Platos problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104(2), 211–240 (1997)
7. Leidner, J.: Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Dissertation.com (2008)
8. Lowe, W.: Towards a theory of semantic space. In: Proceedings of CogSci. pp. 576–581 (2001)
9. Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., Wellner, B.: Spatialml: Annotation scheme, corpora, and tools. In: Proceedings of LREC'08. Marrakech, Morocco (May 2008)
10. Mei, Q., Liu, C., Su, H., Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proceedings of WWW '06. pp. 533–542. ACM, New York, NY, USA (2006)
11. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: Proceedings of ACL. Columbus, OH (2008)
12. Moretti, F.: Atlas of the European Novel 1800-1900. Verso (1999)
13. Overell, S.: Geographic Information Retrieval: Classification, Disambiguation and Modelling. Ph.D. thesis, Imperial College London (2009)
14. Smith, D.A., Crane, G.: Disambiguating geographic names in a historical digital library. In: Proceedings of ECDL'01. pp. 127–136 (2001)

[5] `http://code.google.com/p/textgrounder/`

The Occurrence and Distribution of Spatial Reference Relative to Discourse Relations

Blake Stephen Howald[1]

Georgetown University, Department of Linguistics, ICC 479, 37th and O Streets, NW
Washington, DC 20057-1051
{bsh25}@georgetown.edu

Abstract. I present a descriptive analysis of reference to physical space in the Penn Discourse TreeBank. In particular, I analyze the occurrence of spatial prepositional phrases relative to the discourse relations and semantic senses that hold between two adjacent clauses. The purpose of this investigation is twofold: (1) to better understand how often spatial reference occurs in discourse and (2) to investigate possible relationships between spatial reference and discourse semantics. Overall, the distribution of spatial prepositional phrases and relation-sense pairs are similar. However, statistical evidence suggests that the inclusion of spatial reference in a given clause is independent of the relation-sense of that clause and adjacent clauses. While these results, as applied to the PDTB, indicate the absence of a default pattern of occurrence and discourse semantic function of spatial information, they can nonetheless be extrapolated to provide crucial insights for fully understanding models of spatial representation and interpretation in discourse generally.

Keywords: Spatial Reference, Discourse Relations.

1 Introduction

The semantic and pragmatic functions of discourse relations, which hold between two clauses, contribute to a text's coherence [1]. For example, in the two-line discourse (a) *Lucy is not hungry* (b) *Cati fed her*, (b) is an EXPLANATION for (a) [2]. The inclusion of spatial reference, while accounted for in definitions of discourse relations (e.g., BACKGROUND), is not strictly necessary. However, recent research, grounded in spatial cognitive psychology (e.g., cognitive maps), has suggested that space plays a larger role in discourse structure; in particular, spatial reference organizes narrative discourse into spatially defined groups of events that are temporally linked [3-4]. While this research presents a new analytical perspective, before it can be fully exploited, it is first necessary to better understand what relationships may exist between spatial reference and discourse relations generally.

[1] I would like to thank two anonymous reviewers, my dissertation advisor E. Graham Katz, James Pustejovsky and David Herman for beneficial insights and discussion.

This paper presents the results of a descriptive analysis that evaluates the interface of spatial information and discourse. The particular research question addressed is: Does the occurrence of spatial reference in discourse pattern relative to discourse relations? A negative answer, which is suggested by existing definitions of discourse relations, indicates that spatial reference is independent of discourse relations. An affirmative answer indicates that spatial reference is dependent on (certain) discourse relations. This paper is arranged as follows: Section 2 discusses spatial information (as defined by The Preposition Project [5]), discourse relations (as defined by Penn Discourse TreeBank (PDTB) [6]) and the methodology employed. Section 3 presents the distribution of spatial prepositional phrases relative to discourse relations. Section 4 concludes.

2 Background, Data and Methodology

In this paper, "spatial reference" refers to physical relationships arranged in *figure* and *ground* relationships. For example, *the cup is on the table* locates the figure *the cup* relative to the ground *the table*.[2] A search algorithm was developed to automatically extract 334 different prepositions defined in the Preposition Project [5] (based on a hierarchical network of dictionary entries). 107 of the 334 prepositions have a distinct "spatial" sense. Because prepositions are highly ambiguous (e.g., numerous non-spatial senses), the prepositions extracted from the PDTB were disambiguated by hand.

The PDTB includes annotations of discourse relations in the Penn Treebank II version of the Wall Street Journal (WSJ) corpus [8]. Discourse relations in the PDTB (which hold between pairs of syntactically classified arguments from Penn TreeBank II) ("ArgPairs") are a confluence of connective words, content of the ArgPairs and semantic senses. ArgPairs are either: *Explicit* – a syntactically classified connective word exists in the text (*but*, *and*); *Implicit* – a connective word does not exist in the text but can be inferred; *EntRel* – no relation holds, but the second clause in the ArgPair includes more information about the first clause; *AltLex* – there is no connective word, but a non-connective expression can capture an inferred relation; and *NoRel* – no relation holds. Explicit, Implicit and AltLex ArgPairs co-occur with one of four senses: *Temporal*, *Contingency*, *Comparison* and *Expansion*. The PDTB includes 2159 annotated documents, 40,600 relations and 34,877 senses in total. The overall distribution of the relations and senses in the PDTB provide a baseline of relation-senses. The occurrence or non-occurrence of spatial reference overall, and relative to particular relation-senses and pairs of relation-senses, can then be compared to this baseline to determine relevant (statistically significant) differences and potential patterns.

[2] For sake of brevity, I am restricting the discussion to figure and ground relationships indexed by spatial prepositions [7]. Other sources include motion verbs (*run*, *follow*), deictic verbs (*come*, *go*) and deictic adverbs (*here*, *there*).

3 Results – Distributions and Dependency

200 documents (approximately 10% of the total PDTB), consisting of 5000 relations and 4388 senses, were selected for analysis. If one or both of the arguments in an ArgPair contained one or more spatial prepositions, then these are referred to as Spatial ArgPairs.[3] The occurrence of Spatial ArgPairs is roughly equally distributed between each argument (Arg1 – 54.15%; Arg2 – 45.84%). The average percentage of Spatial ArgPairs per document is 28.90%. The sample selected for analysis conforms to the general relation and sense distributions in the PDTB (Table 1).

Table 1. Distribution of relations and senses.

Relations	PDTB (%)	Sample (%)	Spatial (%)	Senses	PDTB (%)	Sample (%)	Spatial (%)
Explicit	18459 (45.46)	2311 (46.22)	605 (41.86)	**Expansion**	15432 (44.24)	1832 (41.75)	524 (43.73)
Implicit	16053 (39.54)	2002 (40.04)	596 (41.24)	**Contingency**	8016 (22.98)	1005 (22.90)	255 (21.28)
EntRel	5210 (12.83)	578 (11.56)	209 (14.46)	**Comparison**	7634 (21.88)	940 (21.42)	272 (22.70)
AltLex	624 (1.54)	75 (1.50)	22 (1.52)	**Temporal**	3795 (10.88)	611 (13.92)	147 (12.27)
NoRel	254 (.63)	34 (.68)	13 (.89)				
Total	**40600**	**5000**	**1445**	**Total**	**34877**	**4388**	**1198**

There does not seem to be any independent pattern demonstrated by the Spatial, as compared to Non-Spatial, ArgPairs. This is supported by X^2. H_0 is that the occurrence or non-occurrence of spatial reference is independent of a given relation-sense. For the top six relation-senses occurring in the sample (Explicit-Expansion (EE), Explicit-Comparison (EP), ENT, and Implicit-Contingency (IC)), H_0 can be accepted as the *p*-value is greater than .05 and rejected for the Implicit-Expansion (IE) and Explicit-Temporal (ET) relation-senses as the *p*-value is less than .05 (Table 2).

Table 2. X^2 for spatial and non-spatial relation-senses and pairs.

Relation-Sense	Non-Spatial	Spatial	*p*-value	Relation-Sense Paris	Non-Spatial	Spatial	*p*-value
IE	1073	384	.0002	**IE - IE**	223	102	.6499
EE	796	197	.0546	**EE - IE**	163	70	.9507
EP	630	175	.6575	**IE - EE**	163	72	.8568
ENT	595	180	.5580	**IE - EP**	128	52	.6953
IC	513	157	.5291	**EP - IE**	118	56	.5738
ET	451	99	.0131	**EE - EE**	110	40	.3421

[3] 40 of the 107 Preposition Project prepositions are represented in the analyzed sample (N = 2214) with common prepositions making up the majority (82.92%): *in* – 880 (39.74%); *at* – 335 (15.13%); *to* – 250 (11.29%); *on* – 142 (6.41%); *from* – 130 (5.07%); *of* – 117 (5.28%). The remaining 36 prepositions account for the 17.08% complement.

However, the effect that is being exhibited by the IE and ET relation-senses arguably has more to do with the occurrence of Non-Spatial ArgPairs because of the comparative number (1073 Spatial vs. 384 Non-Spatial for IE and 796 vs. 197 for EE). For pairs of relation-sense s, H_0 can be accepted in all cases (the top six pairs of relation-senses in Table 2) as the p-value is greater than .05. This indicates that, even in greater local context, the occurrence or non-occurrence of spatial information is independent of a given pair of relation-senses.

4 Conclusions and Limitations

In sum, as applied to the PDTB, for the studied sample, there is statistical evidence to support a negative answer to the posed research question: whether or not a figure and ground relationship occurs, indexed by a spatial preposition, is independent of the type of discourse relation. This insight may prove useful in interpreting the results of computational tasks that interpret, represent and analyze spatial information in discourse. The main limitations in this study are the amount of data and scope. Future research will focus on more linguistic spatial phenomenon and larger corpora with varied genres (the WSJ corpus consists of Essays, Summaries, Letters and News; the latter of which accounts for roughly 90% of all text in the corpus [9]). Nonetheless, the present results facilitate a more complete understanding of spatial reference in discourse structure. The occurrence of spatial reference does not appear to be biased by inherent discourse patterning.

References

1. Hobbs, J.: On the Coherence and Structure of Discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University (1985)
2. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press, Cambridge, UK (2003)
3. Herman, D.: Spatial Reference in Narrative Domains. Text 21(4), 515--541 (2001)
4. Howald, B.: Granularity Contours and Event Domain Classifications in Spatially Rich Narratives of Crime. COSIT 2009 Workshop on Presenting Spatial Information: Granularity, Relevance, and Integration, Aber Wrac'h, France, http://repository.unimelb.edu.au/10187/5516 (2009)
5. Litkowski, K.:Digraph analysis of dictionary preposition definitions. In: Proceedings of the SIGLEX/ SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, pp. 9--16. Association for Computational Linguistics (2002)
6. Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., Webber, B.: The Penn Discourse Treebank 2.0 Annotation Manual. The PDTB Research Group (2007)
7. Asbury, A., Gehrke, B., van Riemsdijk, H., Zwarts, J.: Introduction: Syntax and Semantics of Spatial P. In: Asbury, A., Dotlačil, J., Gehrke, B., Nouwen, R. (eds.) Syntax and Semantics of Spatial P, pp. 1--32. John Benjamins, Amsterdam & Philadelphia (2008)
8. Marcus, M., Santorini, B., Marcinkiewicz, M.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics 19(2), 313--330 (1993)
9. Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., Joshi, A.: Easily Identifiable Discourse Relations. COLING 2008 (2008)

Interpreting Spatial Relational Nouns in Japanese

Sumiyo Nishiguchi

Center for International Cooperation in Engineering Education
Graduate School of Advanced Technology and Science
The University of Tokushima
2-1 Minami-josanjima, Tokushima-city, Tokushima 770-8506, Japan
nishiguchi@cicee.tokushima-u.ac.jp

Abstract. This paper discusses spatial terms in Japanese. Japanese is a language that uses common nouns such as *ue* "on/over/above" and *naka* "inside" to represent spatial and temporal locations as opposed to languages like English, which uses prepositions such as *on*, *in*, *under* and *between* to express spatial locations. I consider Japanese common nouns for spatial locations to be relational nouns that are two-place predicates, one of whose argument slots is filled with the entity represented by the other NP in the NP_1-*no* NP_2 construction. The corpus data [1] suggests that spatial nouns are often semantically ambiguous among physical, metaphorical and temporal locations. Therefore, the ontological information used in the Generative Lexicon (GL) [5] is useful for spatial term disambiguation.

1 Spatial Relational Nouns

This study considers common nouns representing spatial locations to be relational nouns. While languages like English use prepositions such as *in*, *on*, *under*, or *between* to represent spatial locations, languages such as Chickasaw in North America use relational nouns to express locations [2].

(1) chokka' pakna'
house top
"the top of the house (the house's roof)"

[2, 4]

Pakna' is a relational noun meaning "top," which follows its possessor *chokka'* "house."

1.1 Spatial Relational Nouns in Japanese

Japanese is one such language that expresses locations by using relational nouns like *naka* "inside," *ue* "on/above," and *shita* "under."

(2) a. mune-no mae-de tenohira-o awase (4179) [1]
chest-GEN front-LOC palms-ACC hold
"Put your palms together in front of the chest"

[1] The numbers in round parentheses indicate the sentence IDs of the output of the data in Yahoo! Chiebukuro section of [1] using ChaKi.NET 1.2β.

b. Sensuikan-no nakat-te kaiteki-desu-ka? (1824)
submarine-GEN inside-TOP comfortable-HON-Q
"Is is comfortable inside of the submarine?"

Mae "front" and *naka* "inside" are relational nouns that do not stand alone semantically. *Mae* "front" is semantically unsaturated so that it always means something's front, as *musuko* "son" is always someone's son, e.g., *Bill's son*. *Mune* "chest" is the argument of the relation represented by *mae* "front." Similarly, *sensuikan* "submarine" is the argument of the relational noun *naka* "inside."

1.2 Non-spatial Relational Nouns

Nouns such as *father*, *friend*, and *enemy* are called relational nouns. Because a father is a father of someone, a friend is of someone, and so is enemy, they are considered to represent functions or relations *father-of*, *friend-of*, and *enemy-of*.

[4] points out that it is the relation expressed by the relational noun *brother* in *John's brother* that the relation between John and his brother inherits, unlike *John's book* in which *book* is a common noun so that the relation between John and his book is not specified—it can mean the book that John owns, wrote or borrowed.

As for Japanese, [3] discusses what he calls *unsaturated nouns* (*hi-howa meishi*) such as *shuyaku* "hero/heroin" of a play, *joshi* "boss" of someone which do not become semantically saturated on their own.

I include what [3] calls unsaturated nouns as relational nouns, e.g., *kazu* "number" in *senpuki-no hane-no kazu* "the number of the blades of a fan," *namae* "name" in (3). Since common nouns are one place holders—a function from individuals to truth values—these relational nouns are two-place holders.

(3) a. $[\![namae\_\text{“}name\text{”}]\!] = \lambda x \lambda y[\text{name-of}(y)(x)]$
b. $[\![shu\,jinko - no\_namae\_\text{“}name\_of\_the\_hero\text{”}]\!] =$
$\lambda x[\text{name-of}(x)(\varepsilon y.\text{hero}(y))]$

1.3 Japanese Spatial Language as Relational Nouns

This paper further regards common nouns that represent spacial locations to be relational nouns. For example, *naka* "inside," *ue* "on/above," and *shita* "under" are two-place holders, and nouns such as *aida* "between" which requires another argument are three-place predicates.

(4) a. $[\![ue\text{“}on/top\text{”}]\!] = \lambda x \lambda y[\text{on}(y)(x)]$
b. $[_{VP}\ [[_{NP}$ kohi-no ue]-ni] [miruku-o] $[_{V}$ ireru]]
coffee-GEN on-DAT milk-ACC put
"put milk on (the surface of) coffee"
c. $[\![kohi - no\_ue\text{“}on\_coffee\text{”}]\!] = \lambda x[\text{on}(\varepsilon y.\text{coffee}(y))(x)]$

(5) a. $[\![aida\text{“}between\text{”}]\!] = \lambda x \lambda y \lambda z[\text{between}(z)(y)(x)]$

b. [$_{PP}$[$_{NP}$[$_{NP}$ ha-to haguki]-no aida]-ni] [$_{vP}$[$_{NP}$ kasu-ga] [$_{VP}$ tamari]] (2908)
teeth-and gum-GEN between-DAT plaques-NOM accumulate
"Plaques accumulate between teeth and gum."

c. $[\![hato\_haguki - no\_aida\_\text{“}the place\_between\_teeth\_and\_gum\text{”}]\!]$ =
λx[between(εy.gum(y))(εz.tooth(z))(x)]

2 Ambiguity among Physical, Metaphorical and Temporal Locations

Spatial Noun	Translation	Instances	Share	Physical Direction(Share)	Metaphor(Share)	Time(Share)
ho	toward	54	0.338	6(0.111)	48(0.889)	
naka	in	34	0.213	21(0.618)	13(0.382)	
aida	between/among	10	0.063	6(0.273)	1(0.1)	3(0.3)
ue	on	9	0.05	5	1	2
mae	in front of/before	6	0.037	5		1
shita	under	6	0.038	6(1)		
ue-no	top	6	0.038		6(1)	
ato	after	4	0.025			4(1)
chikaku	near	4	0.025	4(1)		
mawari	around	3	0.019	3(1)		
shita-no	under	2	0.013		2(1)	
tonari-no	next to	2	0.013	2(1)		
ura	back	2	0.012	2(1)		
atari	around	1	0.006	1(1)		
ato-no	after	1	0.006			1(1)
chokuzen	immediately before	1	0.006	1(1)		
chuo	center	1	0.006	1(1)		
chushin	center	1	0.006		1(1)	
fuchi	edge	1	0.006	1(1)		
gawa	side	1	0.006	1(1)		
ge	low	1	0.006		1(1)	
hidarigawa	to the left side of	1	0.006	1(1)		
mannaka	in the middle of	1	0.006	1(1)		
moto	under	1	0.006		1(1)	
mukogawa	over	1	0.006	1(1)		
omote	surface	1	0.006	1(1)		
sayu	to the both sides of	1	0.006	1(1)		
soba	beside	1	0.006	1(1)		
soto	outside	1	0.006	1(1)		
uragawa	backside	1	0.006	1(1)		
ushiro	behind	1	0.006	1(1)		
yoko	beside	1	0.006	1(1)		
TOTAL		160	1	75	74	11

Table 1. Distribution of Spatial Nouns among 3083 NP1-no NP2 Occurrences in *Yahoo! Chiebukuro* portion of [1]

Table 1 suggests that Japanese relational nouns are ambiguous between three kinds of readings, namely, locational meaning, metaphorical location, and temporal sequence. For example, the word most frequent *ho* "toward" is mostly used for comparisons and show preferences toward the better one as in (6a), rather than being used for physical directions as in (6b).

(6) a. Chunichi-yori Hanshin-no ho-ga tsuyoi (2219)
Chunichi Dragons-than Hanshin Tigers-GEN direction-NOM strong
"Chunichi Dragons is stronger than Hanshin Tigers"

b. (neko-ga) watashi-no ho-e ki-masu. (5177)
cat-NOM me-GEN direction-GOAL come-HON
"Cats come toward me."

Mae "front/before," on the other hand, is ambiguous between physical and temporal locations, e.g., *shuppatsu-no mae* "before departure" (4000) and *mune-no mae* "in front of the chest" (4179).

On the contrary, *ue-no* "TOP-GEN" in the *ue-no NP* construction is unambiguously used metaphorically. *Ato* "after" only applies to temporal order while *ushiro* "back" only implies literal location. Similarly, *ue* "top" is used for physical locations in the *NP-no ue* "on NP" construction. However, abstract nouns cannot use *ue* but form noun compounds with a suffix *jo* "on."

(7) a. netto-jo-de iroiro mite-tara (3508)
internet-on-LOC various watch-then
"While surfing on the internet"
b. netto-no ue-de iroiro mite-tara
internet-GEN on-LOC various watch-then
"While surfing on the internet"

(8) a. kohi-no ue-ni awadate-ta miruku-o funwari ire-ta nomimono (6320)
coffee-GEN on-DAT whip-PAST milk-ACC to float put-PAST drink
"a drink of coffee with whipped cream floating on it"
b. *kohi-jo-ni awadate-ta miruku-o funwari ire-ta nomimono
coffee-on-DAT whip-PAST milk-ACC to float put-PAST drink
"a drink of coffee with whipped cream floating on it"

3 Disambiguation of Spatial Language Using Generative Lexicon

The Generative Lexicon (GL) theory [5] is a powerful tool for disambiguation of spatial terms because it provides richer semantic information to the lexicon. GL incorporates an additional lexical entry to the meaning of words called the qualia structure—constitutive (part-whole relation), formal (ontological categories, shape, color), telic (purpose), and agentive (origin).

The formal quale in GL contains ontological information. For example in (8a), coffee is a drink according to its formal quale, and its higher ontological category is a physical entity, which implies that *ue* "on" is interpreted physically. Furthermore, feature matching between relational nouns the other NP is the key to disambiguation of spatial nouns.

References

1. BCCWJ: Balanced Corpus of Contemporary Written Japanese, BCCWJ2009 edition. The National Institute of Japanese Language (2009)
2. Lillehaugen, B.D., Munro, P.: Prepositions and relational nouns in a typology of component part locatives (2006), `http://www.linguistics.ucla.edu/people/grads/lillehaugen/LillehaugenMunro2006aho.pdf`
3. Nishiyama, Y.: Nihongo Meishiku-no Imiron-to Goyoron: Shijiteki Meishiku-to Hishijiteki Meishiku. Hitsuji Shobo, Tokyo (2003)
4. Partee, B.H.: Genitives: A case study. In: van Benthem, J., ter Meulen, A. (eds.) Handbook of Logic and Language, pp. 464–470. Elsevier, Amsterdam (1983, 1997)
5. Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge (1995)

From Data Collection to Analysis – Exploring Regional Linguistic Variation in Route Directions by Spatially-Stratified Web Sampling

Sen Xu[1], Anuj Jaiswal[2], Xiao Zhang[3], Alexander Klippel[1],
Prasenjit Mitra[2] and Alan MacEachren[1]

[1] GeoVista Center, Department of Geography, Pennsylvania State University, U.S.A.
[2] College of Information Science and Technology, Pennsylvania State University, U.S.A.
[3] Department of Computer Science and Engineering, Pennsylvania State University, U.S.A.

Abstract. How spatial language varies regionally? This study investigates the possibility of exploring regional linguistic variations in spatial language by collecting and analyzing a Spatially-strAtified Route Direction Corpus (SARD Corpus) from volunteered spatial language text on the Web. Because of the fast content sharing functionality of the World Wide Web, it quickly becomes a hotbed for volunteered spatial language text, such as directions on hotels' Websites. These route directions can serve as a representation of everyday spatial language usage on the WWW. The spatial coverage and abundance of the data source is appealing while collecting and analyzing large quantities of spatially distributed data is still challenging. Through automated crawling, classifying and geo-referencing web documents containing route directions from the web, the SARD Corpus has been built covering the U.S., the U.K. and Australia. We implement a semantic categorical analysis scheme to explore regional variations in cardinal versus relative direction usages. Preliminary results show both similarity and differences at national level and geographic patterns at regional level. The design and implementation of building a geo-referenced large-scale corpus from Web documents offers a methodological contribution to corpus linguistics, spatial cognition, and the GISciences.

Keywords: Spatial language analysis, volunteered spatial information, geo-referenced web sampling, regional linguistic variation, cardinal directions

1 Introduction

Spatial language is an important medium through which we study the representation, perception, and communication of spatial information. Research has approached spatial language from various perspectives. From the cognitive perspective, research has focused on group or individual differences, on how language affects way-finding behaviour, or on how regional context affects spatial language usage. From the computational perspective, modelling and reasoning has been applied to spatial language interpretation. The spatial language samples used in these studies have been mostly collected by individuals via time consuming experiments or interviews. This data collection method could provide samples that offer understanding on small-scale phenomenon through manual interpretation by analysts.

However, studying the regional linguistic patterns in spatial language—such as regional variations in route directions—requires a spatially distributed corpus. Spatial language data available from the WWW has great potential for this study because of its unrivaled coverage and easy accessibility. For example, it is common to find hotels, companies and institutions offering route directions on their website which provides spatial way-finding instructions to travelers from different places. Harnessing these human generated route directions on-line and analyzing them is the major focus of this study.

2 Methods

To harness route direction documents from the WWW and ensure the spatial coverage of the resulting corpus, a data collection scheme involving web crawling, text classification, and geo-referencing has been developed. Computational tools have been applied for assisting processing the Spatially-strAtified Route Direction Corpus (SARD Corpus) and interpretation of the results.

Collecting route direction documents from the WWW has two main challenges. First, route directions have a high linguistic complexity that makes it difficult to separate the route direction documents from a variety of irrelevant web documents. This challenge can be solved by applying a machine learning algorithms for text classification [1]. The precision of this route direction document classifier used in this study reaches 93% (from 438 positive classified documents, 407 are hand examined to be spatial language documents). Second, exploring regional variation in spatial language usage requires geo-referencing each document in the corpus, which is not an easy task (i.e., Geographic Named Entity Disambiguation). However, postal code, which commonly appears in destination addresses in route directions, can be used to coarsely geo-reference a route direction document on a postal code level. The data collection scheme first utilizes lists of postal codes for crawling web documents. The returned web documents are fed into the route direction classifier, where only positively classified route direction documents are stored in the result corpus. This data collection scheme maximizes the spatial coverage of the SARD Corpus at a postal code level. To prepare the corpus for extracting region linguistic attributes, the SARD Corpus is organized first by nation, then by region (states in the U.S. and Australia, postal district in the U.K.).

The data analysis of spatial language usage in route directions focuses on the regional linguistic variation, which is addressed by analyzing the semantic usages of cardinal directions (i.e.: *north, south, east, west, northeast, northwest, southeast* and *southwest*) and relative directions (i.e., *left* and *right*). The semantic categories used are detailed in Table 1. The scale and size of the corpus makes corpus linguistic tools a necessity for processing the regional linguistic characteristics. The TermTree tool [2], which is a text processing tool with the capacity to handle regular expressions, is used for assisting an analyst to manually evaluate the semantic usages of direction terms. The semantic categorical data is considered regional linguistic characteristics for each region in the SARD Corpus. Visual Inquiry Toolkit [3] is used for geovisualization of the regional linguistic characteristics (Fig. 3) to interpret the analysis result.

Table 1. Semantic categories for cardinal directions and relative directions.

	Semantic categories	**examples**
Relative Direction	1. **Change of direction**	*take a left, bear right*
	2. **Static spatial relationship**	*see a landmark on your right, the destination is left to a landmark*
	3. Driving aid	*keep to the left lane, merge to the right lane*
Cardinal Direction	1. **Change of direction**	*head north, traveling south*
	2. **Static spatial relationship**	*veer southwest on US Hwy 24, turn north*
	3. Traveling direction	*2 blocks east of landmark*
	4. General origin	*from North, if coming from South of New York*
	*used in POI names	*North Atherton Street, West Street.*

As a result of the data collection, the SARD Corpus has been built with 11,254 web documents covering the U.S., the U.K., and Australia. Overview of the workflow is presented in Fig. 1.

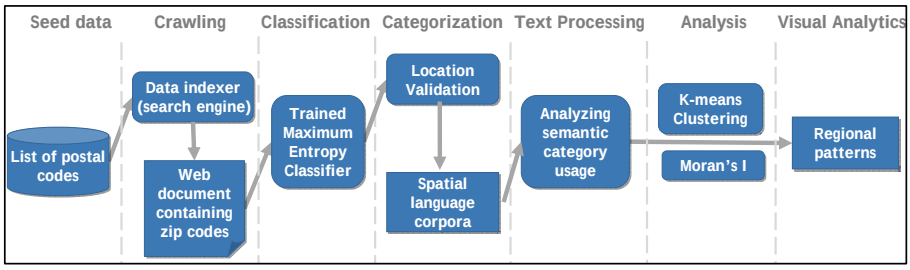

Fig. 1. Overview of the data collection and analysis schemes for building and analyzing the SARD Corpus

3 Results

Regional pattern analysis demonstrates how cardinal/relative directions usage varies at both national level (Fig. 2) and regional level (Fig. 3). On a national level, relative directions in all three nations are mostly used to represent "change of direction" (the blue bar on the left). Similarly cardinal directions are mostly used to represent "travelling direction" (The white bar on the right). On the other hand, the preference for relative direction when representing "change of direction" is much more common in the U.K. than in the U.S. and Australia. Correspondingly we find that cardinal directions are used more often in the U.S. and Australia than in the U.K. (the blue bars on the right) to represent "change of direction".

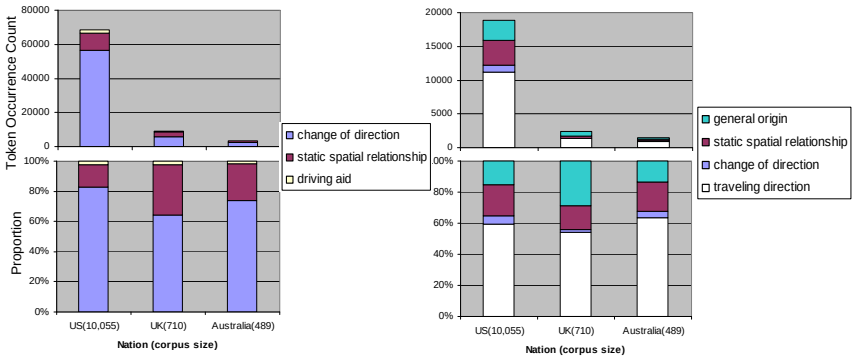

Fig. 2. Nation-level comparison of relative directions and cardinal directions usage

To get a better understanding of the regional variation of relative versus cardinal direction usages, the proportion of each semantic category is plotted on a map for comparison. The plotted map can provide geographical knowledge about the regions, such as adjacency, which helps the analyst to detect regional patterns. Fig. 3 shows that the two most dominant usages as noted at the national-level (relative directions used for "change of direction", cardinal directions used as "travelling direction") are used more frequently in most states in the U.S. For cardinal direction usage, there is a geographic pattern (South Dakota to Kansas, Wyoming to Iowa, blue circle) that differs from its surroundings states in every semantic category. The regional pattern detected is comparable to the Colorado West and Central West region in the map of U.S. dialect [4, p.186]. A possible explanation for this observation may lie in the correlation between the regional linguistic preference and regional geographical features, which is yet to be investigated.

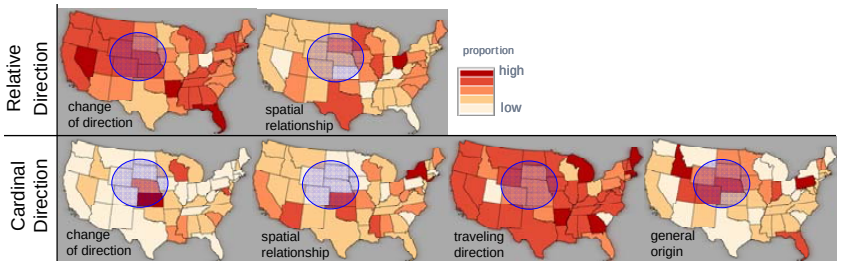

Fig. 3. Regional-level comparison of relative directions and cardinal directions usage (the U.S.).

4 Summary

This paper presents a first step toward an effective and scalable data collection method for spatial language study. It enables spatial cognitive researches to scale-up the spatial language data sets and answer spatial cognitive questions (such as the regional spatial language difference) at a large scale. This study shows promise for effective spatial cognitive research through processing and analyzing volunteered spatial language data, which is an alternative compared to collecting data by designing human participant involved experiments. The presented workflow can also be extended to languages other than English to assist in cross-language comparisons.

The language preference at the nation-level and region-level are both explored, offering 1) a better understanding of how people tend to use spatial language to communicate spatial information; 2) how people differ in using spatial language from different regions; and 3) a guideline to develop a localized, use-specific natural language generation system for navigational devices. Regional patterns of cardinal and relative direction usages in route directions are observed and analyzed, offering a novel perspective for spatial linguistic studies. The design and implementation of building a geo-referenced large-scale corpus from Web documents in this study offers a methodological contribution to corpus linguistics, spatial cognition, and GISciences.

5 Acknowledgement

Research for this paper is based upon work supported National Geospatial-Intelligence Agency/NGA through the NGA University Research Initiative Program/NURI program. The views, opinions, and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the National Geospatial-Intelligence Agency, or the U.S. Government.

References

[1] Zhang, X., Mitra, P., Xu, S., Jaiswal, A.R., Klippel, A., MacEachren, A.M.: Extracting route directions from web pages. In: Twelfth International Workshop on the Web and Databases (WebDB 2009), Providence, Rhode Island, USA. (2009)

[2] Turton, I., MacEachren, A.: Visualizing unstructured text documents using trees and maps. In: GIScience workshop, Park City, Utah (2008)

[3] Chen, J., MacEachren, A.M., Guo, D.: Visual inquiry toolkit - an integrated approach for exploring and interpreting space-time, multivariate patterns. Technical report, GeoVista Center and Department of Geography Pennsylvania State University, Department of Geography University of South Carolina (2007)

[4] Smith, J.: Bum bags and fanny packs: a British-American, American-British dictionary. Carroll & Graf Publishers (2006)

Spatial Relations for Positioning Objects in a Cabinet

Yohei Kurata[1] and Hui Shi[2]

[1] Department of Tourism Sciences, Tokyo Metropolitan University
1-1 Minami-Osawa, Hachioji, Tokyo 192-0397, Japan
ykurata@tmu.ac.jp
[2] SFB-TR/8 Spatial Cognition SFB/TR8 Spatial Cognition, Universität Bremen
Postfach 330 440, 28334 Bremen, Germany
shi@informatik.uni-bremen.de

Abstract. This paper proposes a set of qualitative spatial relations designed for supporting human-machine communication about objects' locations in 'planar' storage. Based on Allen's interval relations and RCC-5 relations, our relations are derived by combining directional and mereo-topological relations between the projections of objects onto the 2D background. We identify 29 realizable relations, which are then mapped to positioning expressions in English.

Keywords: spatial communication, relative locations, mereo-topological relations, directional relations, RCC-5, Allen's interval relations

1 Introduction

When people describe the location of an object, they often use the relative position of the object with respect to other objects which can be identified more easily. In order that people can communicate with smart environment via natural dialogue, computers should be able to understand and generate such positioning expressions. To process such positioning expressions, we may apply existing models of cardinal directional relations (e.g., [1, 2]) or those of mereo-topological relations (e.g., [3, 4]). However, existing direction models distinguish too large number of relations—for instance, Papadias and Sellis [1], Cicerone and Felice [5], and Kurata and Shi [6] distinguish 169, 218, and 222 relations, respectively. In addition, we have to care the application of mereo-topological relations, because *partonomy* actually does not hold between physical objects. This paper, therefore, proposes a task-oriented set of qualitative spatial relations designed for supporting human-machine communication about object locations in a *cabinet*, based on the model of cardinal directional relations in [1] and that of mereo-topological relations in [3] (Section 2). Here a cabinet refers to any 'planar' storage in which we can neglect the front-back arrangement of two different objects. Moreover, objects are limited to *physical* objects in the real world (i.e., 3D single-component spatial objects without cuts or spikes, which never intersect with each other). The resulting relations, called *cabinet relations*, are smoothly mapped to natural language expressions for positioning objects (Section 3).

2 Formalization of Cabinet Relations

Allen [7] distinguished 13 relations between two intervals. Considering projections of 2D objects onto *x*- and *y*-axes and the interval relations between these projections on each axis, Guesgen [8] distinguished 13×13 relations between two 2D objects. His theory, called *Rectangular Algebra* (*RA*), is used typically for capturing *north-south-east-west* relationships, but here we use it for capturing *above-below-left-right* relationships in a cabinet, setting *x*-, *y*-, and *z*-axes parallel to the cabinet's width, height, and depth axes and considering the projections of 3D objects onto the *xy*-plane. Moreover, we summarize the 13 interval relations into 6 relations (Fig. 1b), such that (i) each relation captures how the main bodies of two intervals overlap and (ii) converse of each relation is uniquely determined. The original 13×13 relations and the new 6×6 relations are called *RA relations* and *simplified RA relations*, respectively. For instance, the arrangement of two objects in Fig. 1a is represented by a RA relation (*meets*, *starts*) or by a simplified RA relation (*proceeds*, *within*).

RCC-5 relations [3] consist of five mereo-topological relations, namely *DR* (*discrete*), *PO* (*partial overlap*), *PP* (*proper part*), *PPI* (*proper part inverse*), and *EQ* (*equal*). In our cabinet scenario, we consider the projection of each object onto *xy*-plane, whose *inner* spaces (i.e., empty spaces enclosed by the projection), if they exist, are filled. Then, considering RCC-5 relations between the space-filled projections, we distinguish three spatial relations between the original objects, namely *separate*, *enclosed*, and *encloses* (Fig. 1c). These three relations capture whether one object is enclosed by another object as seen from the front of the cabinet, thereby called *enclosure relations*. Note that the projections never take *PO* and *EQ* relations, since in our scenario two objects never overlap nor have a front-back arrangement.

A *cabinet relation* between two objects is defined as a pair of their enclosure relation and simplified RA relation. For instance, the cabinet relation in Fig. 1a is represented as [*separate*, (*proceeds*, *within*)]. Since we have 3 enclosure relations and 6×6 simplified RA relations, there are 3×6×6 = 108 pairs of relations. However, only 29 pairs (Fig. 2) are realizable in the real world because (i) when the enclosure relation is *enclosed*, the simplified RA relation must be (*within*, *within*) (note that (*within*, *equal*), (*equal*, *within*), and (*equal*, *equal*) are impossible because two objects never overlap nor have a front-back arrangement), (ii) similarly, when the enclosure relation is *encloses*, the simplified RA relation must be must be (*includes*, *includes*), and (iii) when the enclosure relation is *separate*, the simplified RA relation can be any but neither (*within*, *includes*), (*within*, *equal*), (*includes*, *within*), (*includes*, *equal*), (*equal*, *within*), (*equal*, *includes*), (*equal*, *equal*), (*equal*, *overlap*), nor (*overlap*, *equal*), since these relations presume the overlap of two objects.

3 Mapping from Cabinet Relations to Positioning Expressions

When people explain the location of an object, they often rely on topological relations between the object and other related object (especially if they intersect) or directional relations between them (especially if they are located separately). Thus, the cabinet

relations, which capture both topological and directional characteristics of objects' arrangements, have certain correspondences to positioning expressions. Indeed, we can map the cabinet relations to the following English expressions:

- [*enclosed*, (*within*, *within*)]→ *A is in B* (Fig. 2a)
- [*encloses*, (*includes*, *includes*)]→ *A contains B* (Fig. 2b)
- [*separate*, (*proceeds*, *proceeds*)] → *A is at the lower left of B* (Fig. 2c)
- [*separate*, (*proceeds*, *within*/*includes*/*equal*/*overlap*)] → *A is at the left of B* (Figs. 2e-h)
- [*separate*, (*within*/*includes*/*equal*/*overlap*, *proceeds*)] → *A is below B* (Figs. 2o, 2s, 2w, and 2y)
- [*separate*, (*within*, *within*)] → *A is surrounded by B* (Fig. 2q)
- [*separate*, (*includes*, *includes*)] → *A surrounds B* (Fig. 2u)

Among 29 cabinet relations, 24 relations are assigned each to a certain expression. Other 5 relations (Figs. 2r, 2v, 2α-2χ) refer to rather complicated arrangements and are difficult to characterize with simple expressions.

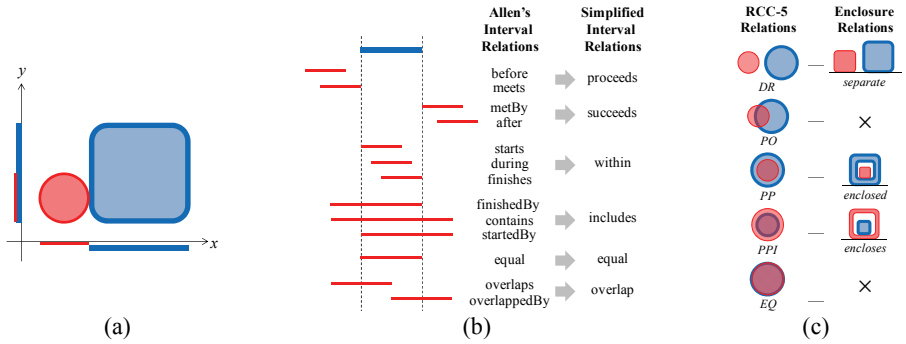

Fig. 1. (a) Projection of two 2D objects in Rectangular Algebra, (b) simplification of Allen's interval relations, and (c) correspondences between RCC-5 relations and enclosure relation

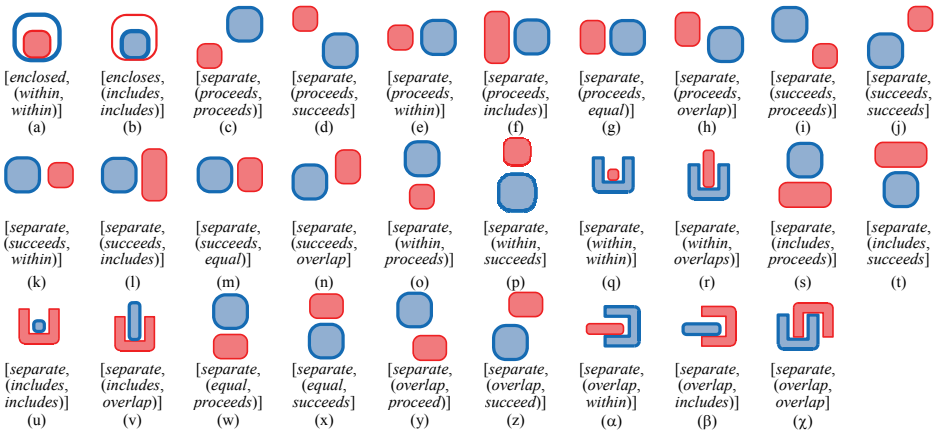

Fig. 2. Twenty-nine cabinet relations

In actual dialogues, people use lots of expressions for describing locations. For generality, we can consider an intermediate use of ontologies. For instance, we can assign [*separate*, (*proceeds*, *within*)] to an ontological concept, which is then mapped to such expressions as "*at the left of*" in English and "*-no hidari-ni*" in Japanese. As a similar work, Shi and Kurata [9] mapped path-landmark relations to ontological concepts in GUM [10]. Such generalization in our model is left for future work.

4 Conclusions and Future Work

This paper introduced a set of qualitative spatial relations designed for the positioning of physical objects in a cabinet. These cabinet relations will work powerfully for supporting human-machine communication in smart environments. At this moment, the mapping between the cabinet relations and language expressions is empirical and thus, we need certain justification of this mapping in future work. We may also need certain fine-tuning of the model, considering the use of additional information such as adjacency/distance between two objects. Lastly, another issue in our future agenda is to implement the proposed idea and test its applicability in practical systems.

References

1. Papadias, D., Sellis, T.: Spatial Reasoning Using Symbolic Arrays. In: Frank, A., Campari, I., Formentini, U. (eds.): International Conference GIS (1992)
2. Goyal, R., Egenhofer, M.: Consistent Queries over Cardinal Directions across Different Levels of Detail. In: Tjoa, A.M., Wagner, R., Al-Zobaidie, A. (eds.): 11th International Workshop on Database and Expert Systems Applications, pp. 876-880 (2000)
3. Randell, D., Cui, Z., Cohn, A.: A Spatial Logic Based on Regions and Connection. In: Nebel, B., Rich, C., Swarout, W. (eds.): Knowledge Representation and Reasoning, pp. 165-176. Morgan Kaufmann, San Francisco, CA, USA (1992)
4. Egenhofer, M., Herring, J.: Categorizing Binary Topological Relationships between Regions, Lines and Points in Geographic Databases. In: Egenhofer, M., Herring, J., Smith, T., Park, K. (eds.): Ncgia Technical Reports 91-7. NCGIA, Santa Barbara, CA, USA (1991)
5. Cicerone, S., Felice, P.: Cardinal Directions between Spatial Objects: The Pairwise-Consistency Problem. Information Science 164, 165-188 (2004)
6. Kurata, Y., Shi, H.: Toward Heterogeneous Cardinal Direction Calculus: . In: Mertsching, B., Hund, M., Aziz, M. (eds.): KI 2009, Lecture Notes in Computer Science, vol. 5803, pp. 452-459. Springer, Berlin/Heidelberg (2009)
7. Allen, J.: Maintaining Knowledge About Temporal Intervals. Communications of the ACM 26, 832-843 (1983)
8. Guesgen, H.: Spatial Reasoning Based on Allen's Temporal Logic. Technical report, International Computer Science Institute (1989)
9. Shi, H., Kurata, Y.: Modeling Ontological Concepts of Motions with Two Projection-Based Spatial Models. In: Gottfried, B., Aghajan , H. (eds.): Behavioral Monitoring and Interpretation, CEUR Workshop Proceedings, vol. 396, pp. 42-56. CEUR-WS.org (2008)
10. Bateman, J., Hois, J., Ross, R., Farrar, S.: The Generalized Upper Model 3.0: Documentation. Technical report, Collaborative Research Center for Spatial Cognition, University of Bremen, Bremen, Germany (2006)

Iconic Gestures with Spatial Semantics: A Case Study

Elizabeth Hinkelman[1]

[1] Galactic Village Games, Inc., 110 Groton Rd., Westford MA 01886 USA
elizh@galactic-village.com

Abstract. The spontaneous gestures that accompany spoken language are particularly suited to conveying spatial information, yet their briefness, individuality, and lack of conventional linguistic structure impede their integration into NLU systems. The current work characterizes spontaneous size gestures in a manual task corpus, clarifying their form, discourse role and representation as a first step toward incorporating them into NLU systems.

Keywords: gesture, spatial language, knowledge representation.

1 Introduction

When gesture carries the primary load of communication, as in the major sign languages, it develops linguistic properties such as verb subcategorization [1] and lexicalization [2,3]. The spontaneous hand gestures that accompany speech, in contrast, do not show linguistic structure [4]. For this reason, computational research on spontaneous gesture has focused primarily on discourse functions, such as using long range video features to signal repair strategies [5] or shifts in topic [6]. Discrete-valued features extracted from gaze and body orientation have also been used for discourse functions such as signaling grounding. Much of this work emphasizes gesture production rather than recognition [7, 8, 9].

Yet the spontaneous hand gestures that accompany speech are increasingly recognized both as a cognitive aid to the gesturer, and an encoding of meaning [10, 11, 12]. Among the spontaneous gestures that accompany speech, *iconic* gestures are those which present "images of concrete entities and actions"[4]. Iconic gestures have in some cases (though not yet broadly) been shown to be effective in communicating spatial information between discourse participants [4, 11, 13].

The current work pursues the incorporation of spontaneous gesture into NLU systems: much groundwork must be laid. Amid the fluidity and abstractness of spontaneous gesture, we focus on concrete gestures with (relatively) straightforward spatial interpretations. We seek to answer the questions:

- What is the discourse purpose of the gestures?
- Do the gestures constitute intended communication?
- To what extent are they lexicalized?
- What are their semantics?
- How can they be related to the semantics of the co-ocurring speech?

2 Corpus study

We collected a reference corpus for dialogue with intonation and gesture in a physical task context. The subjects were twelve pairs of University of Chicago undergraduate and graduate students, who were familiar with each other and had some cooking experience. They were recorded while performing a 30-45 minute cooking task (making chocolate truffles), using a single camera and lapel microphones. Some elements of the task include locating ingredients and equipment, dividing the labor, choosing flavorings, and activities such as measuring and washing up.

The resulting eight hours of videotape were examined for spatial gestures. These included pointing, displaying, miming of physical actions and manner[14], and size gestures. We selected the size gestures as a focus for possible NLU because they are the simplest and most imagistic of these groupings, and because they were relatively uniform in form.

All of the size gestures in our corpus stemmed from the recipe step: "Take a hunk of set ganache and roll into a walnut-sized ball between your palms." An example can be seen in Illustration 1, where subject Chris reads the recipe step aloud, envisions the ball he will roll, and enlists Jason to confirm the ball size. In total he performs the gesture for about three seconds; Jason eventually turns his head to view it for about 800ms. We will refer to this example and similar gestures as 'the ball size gesture'.

2.1 results: ball size gesture use and discourse purpose

Of twelve pairs of subjects, two did not communicate about truffle size beyond reading the recipe. Ten discussed truffle size verbally; of these, three did not use gestures, and three used displays of ganache (dough). Four used size gestures: three ball size gestures and one caliper size gesture[1]. Gestures were used in two main ways: to inform the partner of a desired size, or to request confirmation that a size was correct. In one case, multiple ball gestures were used to explain how an incorrect ball size leads to difficulties in baking. All gestures were used with co-occurring speech.

1 A 'caliper gesture' shows the size of a small object using parallel thumb and forefinger .

2.2 Intended communication – ball size and display

We classify five of the seven gestures as intended communication, on the basis that: in three cases the gesturer used motion or location to attract visual attention; in two cases the gesturer made a verbal reference to the gesture (e.g."like this?"), and in one case both were used. For the seventh gesture (the incorrect ball size explanation) we have no evidence that the gesture per se was intended communicatively. A further analysis of gaze and uptake in these cases is in progress. Although this is a very small sample, most of these gestures showed evidence of communicative intent.

2.3 Form constraints on the ball size gesture

We initially suspected that the ball size gesture was strongly lexicalized in comparison with spontaneous gesture generally. In all cases the thumb and forefinger circle to touch each other and embrace a notional ball, and are displayed as the focal side of the gesture. However, there is notable variation in other parameters. Either hand could be used, as in ASL. The position of the other three fingers is not conventionalized (where it might or might not be constrained in a sign language.) The location of the gesture relative to the gesturer is not as conventionalized as it would be in ASL. In the table, we refer to the gesturer as G and the observer as O.

The third column, the explanation of how two balls may melt into each other while baking, is more typical of spontaneous gesture in showing dynamic configurational elements with extended duration. The ball size gesture is not as conventionalized as an ASL gesture – nor can we say what lexicon it would belong to. More work is needed on this point. The ball size gesture contrasts with the caliper gesture in form.

Lexicalized?	Chris&Jason	Chris&Trish	Josh&Naomi
Hand	left	right	both
Handform	'OK'	'OK'	'OK', 'OK'
Fingers	splayed	curled	splayed, splayed
Orientation	O's visual plane	O's visual plane	Off G's vis plane
Location	At G's eye level	Near O's focus	Near G's chest
Path	static	static	Slowly together
Duration (ASL=250ms)	>3000ms (G) > 700ms (O)	260ms	1500ms

3 Representing Size

Finally we consider semantic representation. A size is a property of a physical object, generally represented as a value on a scale, where a scale is a partial ordering on a set of elements. The majority of verbal size descriptions followed the recipe text: 'the

size of a" small object, or simply mentioned a small object: walnut, half a walnut, meatball. The comparative "...smaller", and (negated) intensifier "don't make it too big!" also occurred. The scale in this case seems to be based on the generics (types) of ball shaped food items, and the asserted relation is purely qualitative. Qualitative representations [15, 16] may prove extensible. Gesture's spatial medium, by contrast, is continuous rather than discrete; the underlying scale is tied to the visual or perhaps kinesic system. What representation could plausibly be generated by the visual system? Our preliminary work investigates low level features in the spirit of [17, 18].

Acknowledgments. This work was supported in part by NSF grant no. IRI-9109914. K-E. McCullough, C. Sidner and R. Jacobs provided valuable discussion.

References

1. Supalla, T.: Serial verbs of motion in American Sign Language. In S. Fischer (Ed.), Theoretical Issues in Sign Language Research. University of Chicago Press (1990)
2. Hoiting, N., Slobin, D.: From Gestures to Signs in the Acquisition of Sign Language. In Duncan, S. D., Cassell, J., Levy, E. T. (Eds.), Gesture and the Dynamic Dimension of Language, pp. 51 - 66. John Benjamins Publishing Company, Philadelphia (2007)
3. Goldin-Meadow, S. Gesture with Speech and Without It. In Duncan Cassell Levy, pp 31-50.
4. McNeill, D.(Ed.), Language and Gesture, pp.2-7. Cambridge Univ. Press, New York (2000).
5. Chen, L., Harper, M., Quek, F.: Gesture Patterns during Speech Repairs. In Proc. icmi, pp.155- Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02), (2002)
6. Eisenstein, J., Barzilay, R., and Davis, R. 2008. Discourse topic and gestural form. In Cohn, A. (Ed.): Proceedings of the 23rd NCAI, pp. 836-841. AAAI Press (2008)
7. Cassell, J., Nakano, Y.I., Bickmore, T.W., Sidner, C., Rich, C.: Non-verbal cues for discourse structure, Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, p.114-123. Toulouse (2001)
8. Traum, D., Morency, L-P.: Integration of Visual Perception in Dialogue Understanding for Virtual Humans in Multi-Party Interaction. In Proc. AAMAS (in press). Toronto (2010)
9. Rich, C., Ponsler, B., Holroyd, A., Sidner, C.L.: Recognizing Engagement in Human-Robot Interaction. In: Proc. Human-robot Interaction. Osaka (2010)
10. McNeill, D.: Gesture and Thought. University of Chicago Press, Chicago (2005)
11. Tversky, B., Lozano, S. C.: Gestures aid both communicators and recipients. In K. Coventry, J. Bateman, T. Tenbrink (Eds.), Spatial language and dialogue. Oxford: Oxford University Press (forthcoming)
12. Goldin-Meadow, S. Hearing gesture: How our hands help us think. Cambridge, MA: Harvard University Press (2003)
13. Beattie, G., Shovelton, H.: When Size Really Matters. Gesture, 6:1., pp. 63-84 (2006)
14. Hinkelman, E.: Spatiomotor Routines as Spontaneous Gestures. Spatial Cognition (2010)
15. Lovett, A., Forbus, K.: Shape is like Space: Modeling Shape Representation as a Set of Qualitative Spatial Relations. AAAI Spring Symposium Series, North America, Mar. 2010.
16. Bateman, J.A., Hois, J., Ross, R. J., Tenbrink, T. A Linguistic Ontology of Space for Natural Language Processing. In Artificial Intelligence, in press (2010)
17. Regier, T., Carlson, L.A.: Grounding Spatial Language in Perception: An Empirical and Computational Investigation. Journal of Experimental Psychology, Vol. 130, No. 2, pp 273-298 (2001)
18. Franconieri, S.L., Scimeca, J.M., Roth, J.C., Helseth, S.A.: Visual Spatial Relationship Representation as a sequence of attentional shifts. Subm. J. Cognitive Science.

Author Index