23

Meaningful Models for Information Access Systems

JUSSI KARLGREN

23.1 Distributional models of language

Study of semantics has the general goal of modeling human linguistic competence as a theory, probing the constraints and limitations of language as a system of expression and representation, and of providing language engineering applications with a model of meaning, appropriate to its tasks. In general, there is no need to design a semantic model intended for practical processing to be neurologically or psychologically plausible but since human performance is impressive in certain respects there certainly is reason to investigate it to find if it can provide inspiration, examples, or constraints for implementations. Human information processing is efficient and effortless. The human information processor is flexible, dynamic, ever learning, does not stumble at inconsistencies, and does not require formal or explicit instruction.

What sort of demands would we want to pose on a model of meaning, from the standpoint of language engineering for information access? Some specific requirements are at the forefront for information access analysis. Information access involves matching brief or even incomplete expressions of information need to relatively more verbose documents and items of information. The documents are not necessarily formulated for ease of retrieval in mind.

For this class of tasks, models that are based on dynamically observed data of language use in some form are dominant. They have common characteristics, however those data are collected and whatever the character of the data: they are based on occurences of linguistic units in a context of use; they do not rely on explicitly represented pre-compiled knowledge; they are flexible

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

and sensitive to the domain and universe of discourse at hand.

The *Distributional Hypothesis*, the basis for distributional language models, states that two words are similar to the extent that they share contexts Harris (1968), and thus that distributional data — of how words appear in contexts — can be used to model similarity, however it is understood, between words. That statement can be used as a basis for a theory of meaning suitable for practical deployment in contexts where approximative semantic analysis of large amounts of linguistic data is necessary, approximating similarity in use with similarity in meaning.

Change or semantic drift is modelled seamlessly by distributional models. New data will provide new occurrence data for the model. The problem of modeling change can be formulated as the problem of selecting the right training context: what data are relevant to the model at hand? If the correct situational context is provided for the model, the resulting representation will reflect the usage in them. This is a desirable quality in the models: we know human language changes fluidly. From one intellectual context to another and from one discourse situation to another the usage and prototypical referents of expressions shift and change with little or no confusion for human users; as time passes, words' meanings evolve and change with little or no confusion, without any attention from their users.

Most distributional models are difficult to provide with precomputed data — to "teach" — in a non-arbitrary manner. Again, this is a desirable quality. We know people learn language their entire life. They do this *without explicit acts of definition and instruction*. In keeping with this it would be useful to find that a system for processing large amounts of text from varying sources have a semantic model capable of operation with little human intervention, with the necessary knowledge extracted from the data at hand. Distributional models in practice are implemented not only to work without supervision but in fact most often to forswear it entirely.

Most distributional models do not rely on external fixed knowledge sources to any great extent, and base their deliberations on statistical or probabilistic calculation on the data alone. We know people seldom take recourse in definitions or formal delimitations of meaning between types of expression. Expressions can be more or less similar in meaning, changing with author and reader perspective or situational context: a semantic model for robust processing of information from many authors to many readers must not be brittle and dependent on exact expression of formal knowledge — it should seamlessly incorporate the gradual shift in meaning from same to similar and from related to distinct (Karlgren, 1976, e.g.). Distributional models are typically implemented with calculation frameworks with intrinsic provision of gradual shades of *homeosemy* or relative similarity.

As can be inferred from the sketchy description above, both *word* or *term*

on the one hand and *context* on the other are central for modeling distributional data. The data may be preprocessed to identify graphical word occurrences, morphologically normalized words, multi-word terms, or whatever linguistic unit is being considered. The nature of the context studied varies according to what sort of model is being built: an utterance, a window of a few surrounding word tokens, an entire text, or a topical unit.

23.2 Representing distributional data — understanding language models

Distributional models collect data of term occurrences. These data are compiled in some representation for convenient further processing. *Probabilistic language models*, e.g., refine the occurrence data into an estimate of the probability that a given word will appear again, given some observed or observable context.

The dominant language model for analysis of textual information in information access and lexicographical applications is the *vector space model*. A vector space is a many-dimensional space where the points can be accessed by address – by a vector of coordinates using some system, typically cartesian. A point in a vector space can be described by a vector $\vec{v}$ thus:

$$\vec{v} = [v_1, \cdots, v_n]$$

where n is the dimensionality of the vector space.

The vector space model for languages posits such a many-dimensional space for terms by populating a vector space with distributional data of term usage in text or discourse. The data are represented in a matrix F of order $w \times n$, such that the rows F_w represent the terms, the columns F_n represent the contexts under consideration — documents, e.g., in the most typical case — and the cells are the (possibly weighted and normalized) frequency of a given term in a given context. Each row of frequency counts thus constitutes an n-dimensional occurrence vector $\vec{v}$ for a given term. These occurrence vectors, interpreted as coordinates in an n-dimensional space as above, deliver a vector space model with the occurrence vector defining a location for its term.

Vector space models have gained increasing currency for application to information access tasks. They exhibit several attractive qualities, not the least being that of pleasing intuitive simplicity, transparency and ease of explication. They are also computationally efficient in several respects, and have proven useful in several applications.

This model lends itself naturally to the application of standard distance metrics. Position is determined by the occurrence of terms in contexts; closeness in space implies distributional similarity or similar usage; and proximity between points — terms — in this space can easily be understood as simi-

$$d_{cos}(\vec{u},\vec{v}) = \frac{\vec{v}\cdot\vec{u}}{\mid \vec{u} \mid\mid \vec{v} \mid} = \frac{\sum_{i=1}^{n} v_i u_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

FIGURE 1 Computation of cosine between two vectors

larity in meaning. This notion of proximity or distance can be used to model gradual shades of relative similarity.

Similarity can be established either by calculating the distance between the points in space, or by transforming the vectors to polar coordinates and using the angle between them. This, in essence, normalizes the relative magnitude of the cell values in the matrix – vectors with the same orientation are considered equal. Most often the cosine of the angle as per the formula in Figure 1 is used: it interprets readily as a proximity measure.

In summary, vector space models localize terms at points in space. Proximity of a term to other terms is calculated through some distance measure. The meaning of a term is found by inspection of its closest neighbors — meaning is considered to be located in a region around terms. Terms can shift meaning, and this is modeled by moving the term to another point in space.

23.3 Space and meaning

As any model, the vector space model is intended to simplify the notion it is modeling, better to aid processing or understanding the object notion; as any metaphor the space and distance metaphor for meaning mediates experience from one area of human activity to another by conceptual transference.

The space metaphor is powerful and pervasive in human thinking and seems to fit in neatly with intuitions about how meaning comes about. Expressions such as “close in meaning” abound. But what sort of space do people think about when they use spatial expressions to discuss meaning?

While relative distance or proximity seem to be central, neither absolute distance measures nor other spatial relations are normally used. Each semantic comparison we make can be made in terms of proximity — no other relations are simple to make explicit. “Close in meaning.” or “Closer in meaning.” are acceptable statements; “*Slightly above in meaning.”, “*More to the north in meaning.” and “*One metre removed in meaning.” are not. It seems that our conception of meaning as space is limited to something like a limited view of a one-dimensional space.

23.4 Distributional models do not preserve all distributional information

While the distributional models base themselves on occurrences in data, they generalize from those observations, thus ridding themselves of overly specific

information. Probabilistic models sample the data and establish estimates of probable reoccurrence of observed items; vector space models compile the occurrence data into a point in vector space. In both cases, a large amount of distributional information is discarded.

The vector space model is useful and attractive, but does have limitations. Some of them have to do with our understanding of the space metaphor itself: the notion of distance between points leads us to the wrong calculations and an incorrect view of what the space is. While the multi-dimensional space *may* be the correct framework to solve structural problems of the representation, our intuitions risk leading us astray.

The intuitive use of the expressions "conceptual distance" or "close in meaning" does not specify in what way that distance is calculated, nor what topological status the locus of "concept" or "meaning" have; neither does the vector space model require a specific distance measure or definition of meaning. Yet the influence of our intuitions from living in two dimensions of a three-dimensional world via grade school geometry to the vector space calculations have led us to a too constrained view of what can be achieved using the model. This constraint may be inherent in the model, but it may also be a constraint only of the metaphor and our representation of the model. Determining whether the metaphor or the model is the limiting factor is difficult or impossible to do without proper calculation; our intuitions about space and meaning are not the right tools to make informed decisions.

The solar system metaphor of an atom is a parallel case of a representation and a model leading its users to wrong conclusions. The solar system model is seductive in its simplicity and its imaginative qualities. A considerable amount of effort in higher physics classes is spent trying to unlearn the model — which has been useful for gaining the first glimpses and first steps of understanding of subatomic structure, but where each obvious successive generalization is a step in the wrong direction.

23.5 Points, distances, and dimensions

Vector space models localize terms at points in space. Terms can shift meaning, which is evidenced by their occurrence data; these data are accommodated in the model by moving the term to another point in space. Relations to other terms change accordingly, and are evidenced by new distances calculated between them. This simple operation adheres well to our intuitions of how points in space can be manipulated. When modeling some types of observable distinctions in meaning made in human discourse it may well be contested in view of its discarding a considerable amount of information.

The study of vagueness, polysemy, generality, and other types of distributionally evident data would be well accommodated by broadening the scope

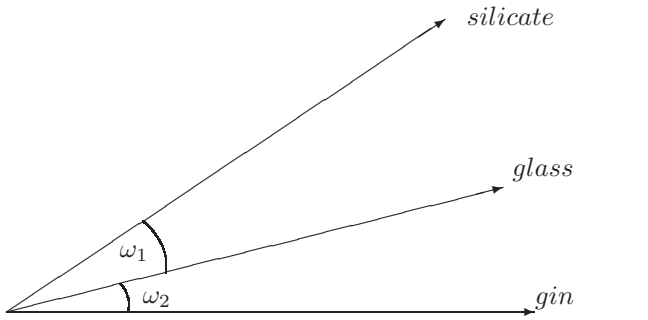


FIGURE 2 Polysemous terms have many kinds of neighbors in two dimensions

of how terms are represented in the model and attendant reform of how the notion of semantic distance is represented.

Distance between two points in a euclidean space is symmetrical and transitively calculable. This does not necessarily always have to be the case in a semantic space. Distance can be calculated in numerous ways. It is possible to examine the implementation of the space metaphor closely, and retool that implementation better to transcend our first intuitions of what geometry is to e.g. allow for non-euclidean, non-symmetric, non-transitive distance measures.

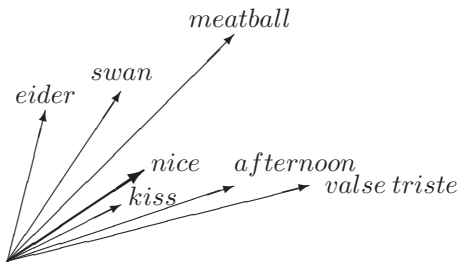
Polysemous terms are a case in point. Proximity between "glass", the beverage, and "gin" on the one hand and between "glass", the substance, and "silicate" on the other need not imply proximity betwen "gin" and "silicate", as illustrated in Figure 2. The risk of confusing transitive proximities can be addressed within the standard term-as-points-framework using additional calculation — by retaining more distributional data in the model and allowing the term to occupy a trace or a more complex structure than a point in vector space.

Vague terms are another example. The capability of vector space models to handle the distinction between vague and definite usage is very limited. If a term in the data is used vaguely, the resulting representation will still try to pull the data together into a point. The representation of a term in the model does not in any way carry the information whether the term should be understood as definite or vague; the distance between terms is calculated identically from a point in vector space whether they are vague or specific. The model pulls together various items as exemplified in Figure 3. It can be argued that the model simply reflects the data: lots of things are nice, and they share a feature. The potential problem with the model is that the vague quality

of niceness is typically modeled as strongly as is the definite quality of, say, animacy or birdness.

In general, measurement of distance can in the given family of vector space models only be calculated between terms — which is of little utility given that the stated objective of most distributional models is to understand the relationship between concepts or whatever notional units of meaning one postulates. A term without a well-defined meaning — arguably the majority of terms — cannot be represented in any other way than as an (typically weighted) average of its occurrences. This distinction, if addressed at all, should be handled on model level. The vector space model does not handle this distinction.

It is not inherently necessary for the model to attempt to fold together the representation of each term into a point. It is a relatively simple extension to investigate terms represented by spaces rather than points, such as clouds, hyperplanes, clusters or concentric structures — it would involve simply imply retaining more data when refining the raw occurrence data and representing the additional data in the vector space. Higher-order distributional characteristics can be utilized to determine which geometry the distribution of a term should be modeled by: patterns of distribution can be modeled by patterns in space rather than using averages, which throw out most of the distributional information. Such an extension, however, will by necessity break the standard metaphor and its distance measure: the distance between two clouds is not well-defined from without the model itself, and needs to be addressed explicitly, not by inheritance via a metaphor.



$$d(kiss, swan) = o(d(kiss, nice) + d(swan, nice))$$

FIGURE 3 A vague term will be close to concrete terms

23.6 More meaningful models?

In conclusion, distributional models in general, and vector space models specifically, risk having their usefulness overshadowed by overly simple metaphors of use which constrain the amount of information extracted from the raw occurrence data upon which they are built. To better accommodate some of the features of the model or to investigate extended calculation bases of the model, higher-order data could be included — e.g. in some of the directions indicated above. By ridding the vector space model from the simple distance metaphor it is delivered with it will lose one of its most appealing qualities – that of pandering to our intuitions – but promises to gain in explicatory power.

23.7 Acknowledgments

The argument above has as its starting points discussions with Magnus Sahlgren, Pentti Kanerva, and Henrik Hållsten. Several valuable points on meaning and its representation are elaborately addressed by Dominic Widdows in a recent and lucid exposé (2005).

References

Harris, Zellig. 1968. *Mathematical Structures of Language*. Interscience publishers.

Karlgren, Hans. 1976. Homeosemy – on the linguistics of information retrieval. In D. E. Walker, H. Karlgren, and M. Kay, eds., *Natural Language in Information Retrieval - Perspectives and Directions for Research*. Stockholm: Skriptor.

Widdows, Dominic. 2005. *Geometry and Meaning*. California: CSLI Publications.