SCALE AND THE CLASSIFICATORY DIMENSION
A Linguistic Approach to Contextual IR

Esben Alfort
Department of International Language Studies and Computational Linguistics
Copenhagen Business School, Copenhagen, Denmark

Keywords: Contextual IR, Ontologies, Subjectivity, Context, Scale, perspectives, Basic level, Individuation.

Abstract: Users of information retrieval systems presumably have (subconscious) reasons for choosing certain ways of formulating their queries. Consequently, the words used may tell us something about the users' intentions. However, researchers in Contextual IR have a strong emphasis on computational solutions and tend to ignore a careful linguistic analysis of the actual queries. As a first step towards such a linguistic treatment, I suggest that we treat classification as a dimension parallel to space and time and learn from our experience with these dimensions in trying to cope with subjectivity in connection with IR systems.

1 RELEVANCE[1]

We are now in the fortunate position of having access to ready-made ontologies for use in semantic applications. It is therefore regrettable that companies producing semantic software cannot benefit fully from these resources, because the resulting ontologies simply contain far too much information. A search engine relying on an elaborate ontology with many kinds of relations would return far too many hits, because most of them would simply not be relevant to a given user in the context at hand. We therefore find ourselves forced to exploit only a fraction of the ontological information available, and even such straightforward relations as subsumption may have to be disregarded in some cases, because the subcategories are judged to be irrelevant to most potential users searching for that particular concept. However, in some cases these subcategories might nonetheless be important. It would thus be preferable to be able to include all the information related to a concept and only exclude some of it from consideration at the last moment if it is judged irrelevant in the case at hand. For instance, when users search for *furniture* on an e-trading site using a semantic search engine based on an ontology, chances are that they are not interested in a complete list of documents on every kind of furniture available. Rather, a link to a page with an overview of furniture in store would be a lot more helpful in most cases. This is because *furniture* is a highly general and heterogeneous category, which does not refer to any specific type of furniture and thus can only be used in generalizations. Consequently, the simple subsumption relation is not very relevant in this case, whereas in other circumstances it may be extremely useful.

Pilot studies of log files from semantic search engines powered by Ankiro (www.ankiro.com) confirm that different user groups characterized by different intentions use search terms from very different classificatory levels (i.e. general vs. specific terms), and that different kinds of sites are visited by different combinations of such user types. Very general categories are characteristic of sites concerned with laws and principles, because rules are generalizations and refer to generalized types. If I were to search for *pets* in a municipal context, I would most likely be looking for laws or suggestions concerning pets in general, and a list of documents on guinea pigs and golden retrievers would be quite unhelpful. Texts on rats would probably be completely irrelevant, dealing mostly with pest control. On a pet shop site, on the other hand, a list of pet types would in all probability be just what I was looking for.

Deciding what is irrelevant is exceedingly tricky, due to the countless facets of relevance (Borlund,

[1] This paper is partly based on a longer version to be presented at a workshop at the Terminology and Knowledge Engineering conference in Dublin 2010.

Alfort E..
SCALE AND THE CLASSIFICATORY DIMENSION - A Linguistic Approach to Contextual IR.
DOI: 10.5220/0003078803700373
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 370-373
ISBN: 978-989-8425-29-4

2003). It seems obvious, though, that we should at least pay close attention to how users actually state their queries. Their linguistic structure may tell us whether they represent a quest for information or a transactional or navigational query (Daoud et al., 2009). To some degree, it may even reveal what kind of information is sought. Unfortunately, this latter possibility is rarely pursued, due to lack of a thorough, linguistic analysis (Tamine-Lechani et al., 2010, p. 6). Instead, the focus has been on developing computational solutions to the problem of identifying the user's context. Whenever the linguistic structure of the query is at all consulted, features noted are rather superficial ones like the number of words in a query, as well as the possible presence of interrogative words in the string (Jansen et al., 2008). Also, certain lexical items that signal a transactional intention (*buy*, *download*, *software*, and so on) are utilized. However, having decided that a request is of the informational type – which is the case in more than 80 % of web search queries according to Jansen et al. (2008) – the actual query is then put to one side, and recourse is in general made to user activity or other sources of information on user interests and preferences. Very frequently, the query is then modified by the addition of specifying and disambiguating terms (e.g. Phinitkar & Sophatsathit, 2010). It is my endeavour to contribute a linguistic view of the matter, in which the query itself plays the main part.

Information retrieval researchers usually look at what documents X are relevant in context C, given a query Q. I shall instead be exploring what expression Q a user would be likely to use in reference to a potential document (or set, or type of documents) X believed to satisfy his or her information need, given the context C. This enhances the understanding of the user's choice of search terms. On the other hand, it masks the fact that users may actually need something different from what they think they do. However, we can hardly solve that problem until we have a satisfactory understanding of the queries themselves.

2 CHOOSING ONE'S WORDS

One is easily led into thinking that every word capable of reference corresponds more or less directly to a certain class of referents. This is in no way the case (Brown, 1958). Even though there are in many cases strong preferences for calling a certain type of entity by one specific name, countless others are possible in principle. When such alternatives are chosen, this is a meaningful act performed by the speaker, and we should understand this as an important piece of information on how the reference is to be interpreted. There are special situations when I might describe a rat by the words *pet, rodent, mammal, animal,* or even *object*. I might also call it a *nuisance*, *friend*, or *test individual*. The sets of potential referents to these expressions differ hugely; consequently, a search containing such terms should differ correspondingly. The focus of my research is currently on why a user chooses a certain hierarchical level in reference to an entity.

I have found that classificatory information can profitably be considered in a way similar to spatial and temporal data, allowing us to apply the understanding that linguists have already acquired of perspectival phenomena in connection with space and time to classification and learn important new lessons about this complex phenomenon. It might not be obvious at first glance how this is possible. I would remind the reader of the fact that the temporal dimension is replete with spatial metaphors like "a short time", "time moves", "we are approaching the future", "from the point of view of fifty years ago", and so on. We do not hesitate in treating the temporal and the spatial dimensions alike, irrespective of whether we consider them to be physically identical or not. This last question is irrelevant here, because ontologies are a cognitive phenomenon, so it is enough that the two phenomena are treated alike in cognition. I propose that we take the next step and recognize classification as yet another dimension. In doing so, I am not suggesting that classification is a physical dimension, but simply that we seem to treat it as such cognitively, and that consequently we can learn much about classification by drawing on our knowledge of spatial and temporal phenomena. Again, I might point to the existence of common metaphors such as "from the point of view of forestry...", in which case forestry is not to be understood as an entity but rather as a topic or mindset, on which all other classification is based (if in any way relevant to forestry). We also speak of "broad topics", "closing in on a topic", and so on, statements that testify to the fact that we think of descriptions (i.e. classification) as something navigable - a plane of sorts.

It would seem that we treat all phenomena in the world that we want to portray linguistically as sets of spatial, temporal and classificatory coordinates.

I shall consider all points regions (cf. Tarski, 1929), since if you zoom in on them they generally turn out to be regions that were simply too small for the internal distinctions to be relevant in the context in question. What corresponds to a region in space or a period of time in classificatory terms is the category - or rather the set of potential referents to which a term applies. The classificatory region is small if a very specific term is used, but includes an increasingly large number of potential referents when more general terms are resorted to.

In portraying situations, we readily zoom in on specific points in space and time that we are concerned about, whereas we generalize and let the distinctions go blurred if we consider them irrelevant in the context at hand. The same is true of classification; we can be highly specific about the type to which a certain referent belongs, or we can get away with very general descriptions if we have no reason to be more explicit.

We can only portray situations as if observed from some point of view. We place an imaginary *observer* at a chosen point in space and time (which I shall call the *host* of the observer) and portray the situation as it appears from that point of view (Alfort, 2009). The host is situated within a *domain of relevance*, which includes the places, times and categories considered, or in other words those regions in all of these dimensions that constitute contrasts to any *focused subregion* of special interest. The phenomenon of *scale* (e.g. Langacker, 2006, p. 116) is caused by the difference in size between the domain of relevance and the focused subregion.

3 CLASSIFICATORY HOMOGENEITY

A referent with a spatial extent may be *homogeneous* if it shows relatively little variation across its spatial region. If we were to zoom in on such a region, we would most likely find that it was not entirely homogeneous, because there are usually small distinctions, which simply pale into insignificance from a larger perspective. However, even greater heterogeneity would be obtained if we were to zoom out, say from a region consisting solely of (a) rock to a section of the seabed on which it was lying. The spatial region would now incorporate other substances such as sand, water and crab. A category may likewise be homogeneous if all potential referents within the domain of interest show relatively little variation. Zooming in on them will make their relative differences grow in importance, and subtypes will emerge. If we zoom out and generalize, however, the category is likely to become even more heterogeneous, as it includes more potential referents of clearly different types.

The world is rarely really homogeneous, so the only way of attaining complete apparent homogeneity is in fact by excluding deviating instances. If I posit that “Italy is a beautiful country”, this is of course a generalization that excludes some less appealing regions such as the occasional refuse tip. Similarly, if someone is “walking in a forest”, he or she is not in fact within anything. Rather, the person in question is walking among the trees that make up the forest that we choose to consider a homogeneous entity for the sake of this statement. In a similar manner, I might say, “If I were a bird, I’d just fly away”, though strictly speaking, if I were an ostrich this would simply not be an option. I thus effectively exclude ostriches from consideration by not including them in the relevant classificatory region behind the term *bird*. I do this in order to be able to treat the category as homogeneous, which is a precondition for a generalization to work - the statement *must* apply to *all relevant* instances. Note that such generalizations are of course impossible in reference to an ostrich, because the homogeneous classificatory region behind the category *bird* would not include ostriches. Consequently, people for whom the ability to fly is a crucial property of birds probably rarely refer to ostriches with the term *bird*, unless the domain of relevance excludes all prototypical kinds of birds, in which case the category would be homogeneous without further generalization. It is for the same reason that we tend not to call boxing gloves *gloves* (Murphy & Lassaline, 1997, p. 110) - such subcategories are simply too deviant to allow generalization across a heterogeneous domain.

4 CLASSIFICATORY LEVELS AND INDIVIDUATION

In connection with space, we are rarely interested in heterogeneous regions; we mostly concern ourselves with the regions that correspond exactly to individuals (at least in connection with prototypical, physical objects). There is a similar tendency to concentrate on homogeneous classificatory regions. This is a phenomenon known traditionally as *basic level* (Rosch et al., 1976). However, I suggest that basic level categories are in fact what one might call *classificatory individuals*, i.e. regions in the classificatory dimension that are relatively homogeneous, and at the same time clearly delimited from surrounding regions. It is important to note that what appears to be an individual is a subjective judgment (cf. Wisniewski et al., 2003, p. 587). When seen from afar, a flock of sheep is considered an individual; it may move about on the hillside, and it may disintegrate into separate sheep, which are also individuals, but as long as there is a flock, it is considered an individual, homogeneous flock of sheep. However, if a shepherd is looking for one particular sheep, he will focus on the individual animals. The flock has become heterogeneous, because different animals are treated as having different coordinates in the spatial dimension. Individuation is a highly subjective phenomenon depending on the granularity of the portrayal as well as the size of the domain of relevance. A vet caring for an injured sheep would focus on that particular animal and would distinguish individual muscles and bones, while the sheep as a whole would be highly heterogeneous for his or her purposes. In the classificatory dimension, the vet would hardly find it useful if all parts of the animal were referred to as instances of *sheep*, since everything within his or her domain of relevance would be sheep. In the same way, the shepherd would hardly refer to separate animals as instances of *flock*, even though they would be, just as drops of water are instances of that liquid.

The fact that individuals are homogeneous regions in space, time and classification means that they dissolve if we zoom too much in on the details, because this makes them heterogeneous. Certainly, individuals have an extremely privileged cognitive status to all humans (Bloom & Kelemen, 1995, p. 7), but this status is not restricted to those entities that appear as individuals in an everyday human context. Rather, whenever we encounter individuals, whether they be flocks, sheep, or muscles, they receive the same privileged status in our consciousness. General and specific terms are used under very different circumstances and for distinct purposes. We can only interpret a query by reference to the context, including the perspective from which the situation is seen in spatial, temporal and classificatory terms. If a very general term is used in a restricted domain of relevance, its meaning is highly dependent on context, because the apparent homogeneity is a product of the restricted domain of relevance rather than an absence of contrast across domains. Compared to this, specific terms are much more straightforward. Consequently, I am currently researching the contexts that allow general terms to be used, and the meanings and intentions behind such expressions. If we can establish the size of the domain of relevance as well as the generality of a term, then this will hint at whether subordinate categories are likely to be relevant to the user providing it.

REFERENCES

Alfort, E., 2009. *Subjektiv og objektiv referenceklassifikation i analyser af kommunikation og tekst.* Master's Thesis, University of Copenhagen.

Bloom, P. & Kelemen, D., 1995. Syntactic cues in the acquisition of collective nouns. *Cognition*, 56, pp. 1-30.

Borlund, P., 2003. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54 (10), pp. 913-925.

Brown, R., 1958. How shall a thing be called? *Psychological Review*, 65, pp. 14-21.

Daoud, M., Tamine-Lechani, L., Duy, D. B. & Boughanem, M., 2009. Towards a graph-based user profile modeling for a session-based personalized search. *Knowledge and Information Systems*, 21, pp. 365-398.

Jansen, B.J., Booth, D.L. & Spink, A., 2008. *Determining the informational, navigational, and transactional intent of Web queries*. *Information Processing and Management*, 44, pp. 1251-1266.

Langacker, R.W., 2006. On the continuous debate about discreteness. *Cognitive Linguistics*, 17 (1), pp. 107-151.

Murphy, G. L., & Lassaline, M. E., 1997. Hierarchical structure in concepts and the basic level of categorization. In K. Lamberts, & D. Shanks (eds), *Knowledge, Concepts, and Categories*, pp. 93-132. Psychology Press.

Phinitkar, P. & Sophatsathit, P., 2010. Personalization of search profile using ant foraging approach. *Lecture Notes in Computer Science 6019, International Conference on Computation Science and Its Applications (ICCSA2010)*, pp. 209-224. Springer Verlag.

Rosch, E., Mervis, C., Gray, W., Johnson, D. & Boyes-Braem, P., 1976. Basic level objects in natural categories. *Cognitive Psychology*, 8, pp. 382-439.

Tamine-Lechani, L., Boughanem, M. & Daoud, M., 2010. Evaluation of contextual information retrieval effectiveness: overview of issues and research. *Knowledge and Information Systems*, 24, pp. 1-34.

Tarski, A., 1929. Foundations of the geometry of solids, reprinted in Tarski, 1982, pp. 24-29.

Tarski, A., 1982. *Logic, Semantics, Metamathematics*. Hackett Publishing Co, 2^{nd} edition.

Wisniewski, E. J., Lamb, C. A., & Middleton, E. L., 2003. On the conceptual basis for the count and mass noun distinction. *Language and Cognitive Processes*, 18 (5/6), pp. 583-624.