IN SEARCH OF A NEW MODEL FOR WSD: A MULTI-LAYERED LEXICAL CONSTELLATION MODEL

Aquilino Sánchez  
Pascual Cantos  
Moisés Almela1  

Universidad de Murcia

ABSTRACT

Research on WSD has been rather intensive in the last quarter of the 20th c. (Agirre & Edmond 2006). Difficulties for a successful disambiguation algorithm lie mainly in polysemous words with high granularity. A research project recently conducted by Sanchez, Cantos, Almela (Sánchez, Cantos & Almela 2007; 2008), grounded on a supervised method, proved that the success rate is clearly lower in words with high granularity, compared to the percentage of success in words with poor semantic granularity. The analysis of the results reveals that failure in the identification of the meaningful contextual features is the main reason leading to disambiguation errors. The analysis concludes that the model of meaning organization used by most WSD algorithms is not adequate to successfully deal with fine-grained sense distinctions. The model of lexical constellation is advanced and sketched as a possibility to initiate a break through in WSD. Its non-linear character and three dimensionality fits better the non-linear and hierarchical structure of meaning organization in real language.

KEY WORDS: WSD, Lexicology, Lexicography, Corpus Linguistics, Computational Linguistics

1. PROBLEMS OF DISAMBIGUATION BY LEXICAL CONTEXT

Studies in corpus linguistics over the last two decades have emphasized the strength of correlations between senses of a word and their lexical contexts (Sinclair 2004). In his Theory of Lexical Priming, which fuses psycholinguistic and corpus-linguistic statements, Hoey (2005) predicts that different senses of a word will be primed to co-occur in different co-textual surroundings. There are thus in the literature strong arguments to support the potential of co-textual data, and collocates in particular, as a sense disambiguating factor. Informed by these theoretical underpinnings, Sánchez, Cantos & Almela elaborated a word sense disambiguation (WSD) algorithm based on the discriminant factor analysis of co-text (See Sánchez, Cantos & Almela, 2007 and 2008 for details on the structure and nature of the algorithm).

Provisional evaluation conducted by members of our research group indicates that the disambiguating power of our algorithm meets the average standards reported by other researchers in the field (Agirre & Edmonds, 2006). Taken as a whole, the results were the ones shown in figure 1 (based on a manually revised sample test of 100 to 150 concordances for each item).

1 This paper derives from a research grant by the Spanish Ministry of Education and Science and by the Murcian Government. The first research project (Ref.: HUM2004-00080/FILO) was funded by the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica. The second project (00481/PI/04) was funded by Fundación Séneca, Comunidad Autónoma de la Región de Murcia.

Contact: Universidad de Murcia, Campus de la Merced, Facultad de Letras, 30071, Murcia, Spain. asanchez@um.es, pcantos@um.es, moisesal@um.es
A closer analysis of the results revealed that the success rate in disambiguating fine-grained senses was lower than expected, particularly if compared to the success rate in disambiguating coarser-grained senses. Our challenge is to discover the flaws leading to failure in our prototype. The analysis of the results of two items (heart and hand) disambiguated by our algorithm will illustrate the case.

The meanings of heart were grouped in 7 senses, as follows:

1. The hollow, muscular organ in a vertebrate animal that receives blood from the veins and pumps it through the arteries by alternate dilation and contraction
2. The human heart considered as the central part or core of something.
3. The human heart considered as the center or source of emotions, personality attributes, etc.,
4. Any of the various humane feelings, like compassion, love, devotion, enthusiasm, etc
5. Something like a heart in shape; conventionalized design or representation of a heart, shaped like this.
6. A red figure of a heart stamped on a playing card, or a playing card marked with a conventionalized figure of a heart.

Structuring the meaning of heart in six senses is somehow a subjective decision. But options are rather limited. Linguists do not agree on what a sense is exactly and dictionaries do not coincide in the amount of the meanings they offer for each entry. The Merriam-Webster Dictionary (2000) specifies 6 main meanings for heart, plus many submeanings within almost each one of them. The meaning of heart as Thoughts / emotions / feelings counts with six sub-meanings. Moreover, some meanings further include more sub-meanings. The New Oxford Dictionary of English (1998) specifies only 4 main meanings for this word. The distribution of meanings and sub-meanings across the semantic field covered is different, both in quantity and content. The sub-meanings for ‘courage, enthusiasm’ and ‘centre of thoughts and emotions’ are included within the basic meaning 1, referred to the ‘muscular organ’, while the meaning ‘the central or innermost part of anything’ –clearly derived from and dependent on meaning 1-, is located at the same level or hierarchy as meaning 1. The same hierarchy is awarded to meaning 3 (‘shape of a heart’). One of the sub-meanings refers to the ‘vital part or essence’ and the other one to ‘the close compact head of a cabbage or lettuce’. We can therefore conclude that the selection taken for our prototype is balanced and reasonable.

We took a sample of 126 concordances of heart from the Lacell corpus and disambiguated them with our algorithm. Meaning 1 was correctly disambiguated in 80% of the instances, while disambiguation of meaning 2 was successful in 39.3% of the sentences, meaning 3 in
49% of the instances, and disambiguation of meaning 4 failed in the two occurrences, as shown in table 1:

<table>
<thead>
<tr>
<th></th>
<th>Total number of samples and %</th>
<th>Success (number of sentences and %)</th>
<th>Failure (number of sentences and %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning 1</td>
<td>50 (39.6%)</td>
<td>40 (80%)</td>
<td>10 (20%)</td>
</tr>
<tr>
<td>Meaning 2</td>
<td>33 (26.1%)</td>
<td>13 (39.3%)</td>
<td>20 (60.7%)</td>
</tr>
<tr>
<td>Meaning 3</td>
<td>41 (32.5%)</td>
<td>20 (48.8%)</td>
<td>21 (51%)</td>
</tr>
<tr>
<td>Meaning 4</td>
<td>2 (1.5%)</td>
<td>0</td>
<td>2 (100%)</td>
</tr>
</tbody>
</table>

Table 1

The percentage of success is fairly high, as a whole. The analysis of errors reveals however that the context taken as a discriminating factor may fail in apparently unexpected instances. The sentence *In this plan he conceived of the Kings Domain as the new cultural HEART of the city where the Shrine would be located attended by the city’s major arts institutions*, was wrongly assigned meaning 1 and meaning 3, having received both the same (rather low) coefficient (0.07662480739740607), while meaning 2, the right one in this case, has not even been selected. Should we assume that the context taken by the algorithm (*Domain as the new cultural HEART of the city where the*) has not discriminated adequately? A human user of the language would probably have gathered enough discriminating information from the word ‘cultural’ and from the group of words ‘of the city’, just taking information from the knowledge of the world humans already have. In fact, the meaning of heart as a ‘physical organ’ does not match with ‘cultural’ or ‘the city’. Our algorithm does not have this kind of world knowledge information and is not even able to include the whole phrase ‘of the city’ –a three word phrase- as a discriminating unit because it processes words individually, unless explicitly told otherwise (as we do for idiomatic expressions). As for ‘cultural’ the contextual information provided to the system has not been ample and rich enough to include this lexical item. The conclusion is that our algorithm fails because it is unable to identify the discriminating information available in the window span considered. The inclusion of ‘cultural’ would ask for enlarging the corpus on which the data base has been built. The failure in the identification of ‘(heart) of the city’ reveals the inadequacy of the algorithm to identify relevant patterns beyond the word level.

2. IN SEARCH OF NEW LINGUISTIC MODELS VALID FOR COMPUTATIONAL PURPOSES

Some researchers have suggested that ‘the traditional WSD research task, so effective at driving research, has reached a plateau’ and ‘is not likely to lead to fundamentally new research’ (Agirre & Edmonds, 2006: 7). They conclude that we need ‘to look for new research directions in the field’. A new feeling is emerging about the need for new models and new computational tools. It seems that the computational tools currently available fail to adequately capture and interpret the necessary and relevant context, particularly beyond the word level. The question is if linguists themselves provide the adequate information on how the context intervenes in sense discrimination and which elements must be taken into consideration in which specific circumstances. While computational experts are responsible for the implementation of linguistic models to be processed by computers, linguists should provide reliable models on how meaning is built along speech or in written text. Traditional linguistic theories are unable to offer any decisive help to break through in WSD. We are in need of models transferable to computational applications, yet flexible and comprehensive to deal with the complexity and variety of semantic realizations as found in natural language.
Sánchez, Cantos & Almela (2007) suggest that results in WSD can be improved with the adoption of the model of lexical constellations, designed by Cantos & Sánchez (2001). This model seeks to solve the problem of how to bridge the gap between the linear surface manifestation of word co-occurrences and the hierarchical organization of lexical structure. If we decide to rely on co-text as the main contributing factor to sense disambiguation, the relevant unit of structural analysis in our WSD algorithm must be the collocate network, rather than the lexical field, which is a purely paradigmatic structure, or the frame, which is primarily a cognitive structure. Collocate networks are made up of groups of words that co-occur significantly with one another, or that share one or more of their significant co-occurrences. Now, one of the specific difficulties in analysing collocate networks is that their display in discourse and corpora is purely linear, horizontal, whereas the semantic bonds and the configuration of conceptual associations among their members is multi-dimensional. This implies that the shared meaning relationships between two or more members of a collocate network will take place within a different semantic dimension than shared meaning relationships with other members of the same network. Besides, the semantic dimensions are often subjected to a hierarchical organization. Thus, the semantic bond between two collocates $a$ and $b$ may imply, in turn, a semantic bond between $(a,b)$ and a third collocate $c$, but conversely, $(a,b)$ is not implied by $c$.

The implication for WSD systems is that correlations between meaning and co-text must be analysed at various levels of granularity. Thus far, decisions concerning granularity in the realm of WSD research have been limited to the selection of lexical entries. There is an ongoing debate about the optimal level of sense granularity in lexical entries for WSD. However, if granularity is a property of sense-context correlations, one and the same mechanism should be able in principle to disambiguate at different levels of granularity, depending on the amount of relevant similarities/dissimilarities observed across uses of a word. Therefore, the potential for oscillating between different levels of sense granularity can be built into the design of the algorithm itself. This requires us to supersede the static treatment of granularity in standard dictionary entries, where senses are further divided into subsenses—or subsenses grouped into broader senses—, and to devise strategies for specifying how a single sentence can be disambiguated at one or other level of granularity. An integral aspect of this strategy is the reorganization of collocates in successive domains and subdomains (lexical constellations), as we illustrate below with the case of the word street.

In some dictionaries, the feature 'outdoor' is recorded as part of a subsense of the noun street. When used with this sense, the word street expresses a contrast between the kind of events that take place inside buildings, houses, etc., and the kind of things that happen out of doors. Compare below the definitions and examples of senses of street in two different dictionaries. The reader will observe that senses 2 and 3 in Macmillan are conflated under sense 2 in Cobuild. However, leaving this subtle divergence aside, the two dictionaries converge on the distinction between the 'road' and the 'outdoor' usage:

1. A road in a town or city with houses or other buildings along it, e.g. Walk along the high street and turn left.
2. Outside in a city or town, e.g. Additional police officers are deployed on the streets.
3. Relating to streets or taking place on the streets in a city or town, e.g. street crime.
   (From Macmillan English Dictionary, 2002)

1. A street is a road in a city, town, or village, usually with houses along it, e.g. Boppard is a small, quaint town with narrow streets.
2. You can use street or streets when talking about activities that happen out of doors in a town rather than inside a building, e.g. Their aim is to raise a million pounds to get the homeless off the streets / ...a New York gang.
In addition to this information from the dictionary entries, it is worth mentioning that the 'outdoor' sense of street has specific socio-geographical connotations, as it is associated with reference to crime in specific urban areas (city centre). The collocations of this word reveal that it "often connotes the inner city and its dangers" Stubbs (2002: 204). These connotations are mutually supported by items from a network of "inter-collocations", as Stubbs calls them, which includes words such as police, gang(s), or youths, among others. The collocational links in this network are distributed as follows (Stubbs 2002: 205):

- streets <walk, police>
- corner(s) <street>
- youths <police, gang(s)>

Let us now analyse this example afresh, in the light of the lexical constellation model. In the network described above, street has three direct collocates (corner, walk, police), one collocate shared with police (gang), and one indirect collocate (youths is attracted towards the co-text of street by two other collocates, namely, police and young). Graphically, the constellation can be represented as shown in Figure 1, where the dotted box represents a sub-domain within the network. The constellation composed of police, gang, and youths, all of which inter-collocate mutually, is tightly knit together around the pivotal notion of 'crime', which is not common to corner or walk. 'Crime' is part of the intrinsic meaning of police, as well as of most of the uses of gang and youths. To a large extent, police activities are oriented towards preventing crime and prosecuting criminals. As for gang and youths, a vast proportion of their uses in the discourse make allusion to groups of criminals or delinquents. Hence, there is in the constellation a subarea (police, gang, youths) which is more densely interconnected by mutual collocation than the rest of the network, and precisely, this subarea represents a specific subdomain in the semantic structure around the node. This reinforces the idea that semantic affinity and syntagmatic attraction correlate at more than one level of specificity.

Figure 1: Constellation of a lexical network around street
What are the implications of this reorganization for the disambiguation of *street*? Note that two collocates of *street* are compatible both with the 'road-sense' (Ex. 1) and with the 'outdoor-sense' (Ex. 2-3). These, so to say, "versatile" or "generic" collocates, which per se do not discriminate between the two main senses of *street*, are precisely those which do not form part of the subdomain (*police, gang, youth*). The latter group of collocates, in contrast, enter a much more specific semantic relationship with *street*. The presence of one or more members of the second group, whether accompanied by a collocate of the first group (as in Ex. 3) or not (Ex. 4), favours the activation of the 'outdoor-sense' of *street*, alongside its socio-geographical connotations of 'crime', 'danger' and 'inner city'. Since the second sense of *street* is a specification the first one, but not conversely, it is in principle possible for the WSD algorithm to deliver simultaneously two different output senses, albeit at different levels of granularity.

Firstly, at the coarser-grained level, the meaning 'road in a city, town, or village, usually with houses along it', is applicable to all the occurrences of *street* in Ex. 1-4. This meaning is extensionally contained within the more specific subsense of 'outdoor', because roads are integral to the settings which witness the criminal activities happening out of doors in the inner part of a town. At this level, the sense-disambiguating potential of collocates such as *police, gang, or youth* is neutralized. The connotative differences between this group of collocates and more neutral ones such as *walk or corner* is considered irrelevant at this stage.

Secondly, at a finer-grained level, the program should be able to detect the differences between Ex. 1, on one side, and Ex. 2-4, on the other side. At this level, the specific socio-geographical connotations triggered by collocates from the sub-domain (*police, gang, youth*) are not to be considered irrelevant. On the contrary, they serve to discriminate between the broad, primary sense of *street* in Ex. 1 and the more specific, connotation-loaded subsense of *street* in 2-4.

(1a) *As you walk* through the *streets* of Funchal, particularly the small, narrow, side *streets*, you can see many artisans practising crafts now rarely seen in the mass-production economies of northern Europe and North America.

(1b) *On the corner* of the *street* is a statue of St Wenceslas by J. J. Bendl of c.1678.

(2a) *I've got to be in the car with the doors locked and, and I just wouldn't walk out in the streets* at night, not because anything has happened to me, but its just through what I've heard, I'm just terrified.

(2b) *Just what we need--400,000 kids hanging out on street corners.*

(3a) Unfortunately this will have to be cancelled as the *police* cannot guarantee the safety of anyone *walking* the *streets* of Manchester, especially anyone that appears to be mincing.

(3b) *As you walk along the street, Puerto Rican youths* whisper offers of Poison; and Cash, two of the brand names of the local heroin, stamped on the bags the drug comes wrapped in.

(3c) *Police rifles were brought out on the corners and in the middle of the streets.*

(3d) Castle ward councillor Henry Spyvee had been particularly concerned about *gangs* of drunken *youths* roaming the *streets* on Saturday afternoons, leaving broken bottles in their wake.

(4a) *Troops continue to patrol the streets of Lhasa and police continue to make arrests - as they do in Peking.*

(4b) *The gang rivalries of the street were kept within bounds and out of the world of adult entertainment.*
Ideally, a WSD system should be able to select the appropriate sense, or bunch of senses, for each of the different levels of sense granularity which can be registered in one and the same lexical entry. One of the main advantages of distributing the disambiguation process in different levels is that, should the disambiguation at finer-grained levels fail, this error will not affect the disambiguation made at coarser-grained levels.

3. Final Remarks

WSD has reached a critical stage: much has been done in the late twenty years or so, and at present we are fully aware of the bottleneck lying ahead of us. Traditional linguistic theories are of little use to solve the problem and computational tools lack a semantic model they might take as a basis for designing automatic disambiguating procedures. The lexical constellation model predicts that correlations of sense and co-text operate at various levels, and it helps WSD algorithms to switch between different sets of output senses depending on the degree of granularity desired. The challenge is then to group collocates into successive domains and sub-domains, basing both on the strength of the links between or among them and on the specificity of the semantic bond with the node. Reaching this goal poses a tremendous pressure on linguists, but it will be difficult to increase the success rate in automatic disambiguation unless the tools used with this purpose are able to manage higher complexity rooted in real and broad, not limited, context.

References