Simonetta Montemagni, Parola sas, Pisa
Stefano Federici, Parola sas, Pisa
Vito Pirrelli, Istituto di Linguistica Computazionale, CNR, Pisa

Example-based Word Sense Disambiguation: a Paradigm-driven Approach*

Abstract

The paper describes an example-based approach to word sense disambiguation: words in a pre-processed text corpus are automatically linked to their corresponding senses in a machine readable dictionary (MRD) by using information automatically extracted from the MRD. For each word sense, typical contexts of use were acquired and structured as "paradigmatic structures" on the basis of distributional criteria. Word sense disambiguation is modelled as a process of "paradigm extension" grounded on the acquired paradigmatic structures. The technique, already applied with success to a number of Natural Language Processing (NLP) applications, is currently under extensive test for word sense disambiguation: preliminary results look promising.

1. Introduction

Over the last ten years, word sense disambiguation has been the focus of increasing attention in NLP circles. Given the occurrence of a polysemous word in a specific context, the task consists in identifying automatically which one of its senses the word is used in. This may not always result in the identification of a uniquely disambiguated sense; in difficult cases, the range of lexical ambiguity can significantly be reduced by isolating a subset of hopefully contextually-relevant senses. An effective approach to this task would be a crucial boost for a number of NLP applications: for example, it would help considerably to reduce recall noise due to polysemy in information retrieval techniques.

In principle, a wide typology of cues can be of avail in choosing among the set of senses defined by an MRD for a given word: they range from syntactic subcategorization to subject domain, to lexico-semantic information (Dolan 1994, Harley forthcoming). The computational procedure described in this paper focuses on the role of lexico-semantic knowledge only. From this it in no way follows that the lexico-semantic constraints which an individual word imposes on its neighbouring words are always sufficient to carry out word sense disambiguation. The algo-

rithm illustrated here can be conceived of as a component of a more complex word sense disambiguation system capable of taking into account the whole range of cues at work in this specific task.

2. Lexico-semantic restrictions and word sense disambiguation

How can lexico-semantic information be used for word sense disambiguation? Consider the case of a polysemous verb: typical argument designations associated with its different senses can be used as disambiguating cues to identify the intended sense of the verb in context. To be more concrete, the Italian verb *accendere* has, according to the Garzanti dictionary (1984), four senses in the transitive reading: 1) to light something which can burn; 2) to switch on an electrical device; 3) to foment a fight or a war; 4) to open a bank account. In all four senses, it is clear that typical objects of the verb play a crucial role as word sense disambiguating cues and are often part and parcel of their definition. The use of typical argument designations for disambiguating the intended sense of a polysemous word is also common practice in bilingual dictionaries to select translation equivalents in the target language. The Collins Italian-English dictionary (1985), for example, says that *accendere* is translated into 1) *light* when the former takes as a direct object nouns like *fiammifero*, *candela*, *sigaretta* (respectively, 'match, candle, cigarette'), 2) *switch on*, when the object is an electrical device such as *radio*, *luce*, *lampada* (respectively, 'radio, light, lamp'), and 3) *open*, if the object is a kind of *conto* 'bank account'.

3. Feature- and example-based approaches to word sense disambiguation

Typically, lexico-semantic constraints imposed by a word on its context are expressed in terms of semantic features: *accendere* in sense 2 of Garzanti can be said to look for an object marked as an [electrical_device]. In practice, however, no lexical resource provides, to our knowledge, feature information of the granularity required for coping with unrestricted texts. Moreover, and most importantly in this context, it is not always the case that the class of words selected, say, as an object of a given verb can easily be expressed in terms of semantic features. Consider the case of Garzanti sense 1 of *accendere*, which requires an object that can be lit (a match, a candle, a cigarette or a fire). If we characterised the class of possible objects of sense 1 of *accendere* in

terms of a general semantic feature – say [physical_entity] – this restriction would not function as a disambiguating cue for sense 1, since radio, light or lamp – i.e. the typical objects of sense 2 of *accendere* – are all marked for the very same feature. For the class containing all and only the objects of sense 1 of *accendere* to be characterised appropriately, one would require very specific and highly ad hoc feature specification, such as, for example, [lightable_physical_entity], with the serious risk that the set of semantic features gets virtually open-ended and their number eventually unmanageable.

Summing up, for sense disambiguation purposes the semantic features expressing the constraints imposed by a word on its context have to be, at the same time, general enough to cover the whole set of collocates and specific enough to rule out spurious interpretations. The balance between these two requirements is not easy to strike and leaves the door open to an explosion of the amount of features required. Example-based approaches (Nagao 1992) have made a special effort to address this issue: semantic features are dispensed with and replaced by actual contexts of use provided by example sentences. The input sentence to disambiguate ("the target sentence") is mapped onto the set of known examples ("example base") in the search of the best analogue, and thesaural information is used to measure the distance between known example sentences and the target. More concretely, an uninterpreted occurrence of – say – *accendere* in the target expression *accendere la televisione* 'switch on the tv' can be assigned the appropriate sense by being mapped onto the known example *accendere la radio* 'switch on the radio', whereby the sense of *accendere* is correctly identified. The mapping is driven by the conceptual relationship between radio and tv, which are both electrical devices and are accordingly specified for the same hyperonym in an available thesaurus.[1]

It should be observed, however, that thesaural relationships (such as hyperonymy, synonymy or meronymy) do not always capture the dimension of similarity which is relevant to the context in which the collocates of a polysemous word are used. Consider the target expression *accendere il carbone* 'light the coal': traditional example-based approaches would arguably pick up *accendere il gas* 'turn on the gas' as the best analogue of the target, on the grounds that both coal and gas are classified as combustible material. The result would be that *accendere il carbone* is wrongly interpreted by analogy to *accendere il gas* (Garzanti and Collins sense 2), and not to the more appropriate *accendere il fuoco*, since the distance between *carbone* and *fuoco* in a thesaurus is bound to be longer than between *carbone* and *gas*.The misinterpretation can be avoided if the semantic similarity between collocates is captured on a

distributional (rather than thesaural) basis by analogy to examples such as *attizzare il fuoco*, *attizzare i carboni* where *carbone* and *fuoco* are distributionally equivalent relative to the same verb sense. In this way, all those collocates which are attested in the example base as subject or object of a given verb sense are considered as somewhat semantically similar. In what follows we illustrate a sense disambiguation system which explores benefits and drawbacks of this paradigm-based strategy.

4. SENSE: basic principles

SENSE (Self-Expanding linguistic kNowledge-base for Sense Elicitation) is an example-based word sense disambiguator which operates on an example base of known verb-noun co-occurrence patterns (where the noun plays the role of either subject or object to the verb) by means of two key notions: paradigmatic structure and paradigm extension (Federici, Pirrelli 1994; Pirrelli, Federici 1994).

4.1 Paradigmatic structures

In the example base, verb-noun patterns are structured on the basis of their actually attested distribution: nouns co-occurring with the same verb sense and playing the same syntactic function (either as a subject or object of the verb) are grouped in sets of semantically similar nouns, called **paradigms**. Here we will neglect the nature of the semantic similarity underlying each paradigm which varies from case to case. Rather, we focus on the way these paradigms are structured and mutually related by SENSE, for them to eventually be used for word sense disambiguation.

Typical instances of the co-occurrence patterns contained in our example base are reported in the table below:

(1)

ABBANDONARE 'abandon'							
paradigm A		paradigm B		paradigm C		paradigm D	
0_1	FAMIGLIA/O 'family'	0_3	CASA/O 'home'	0_4	STUDIO/O 'study'	0_5	REDINE/O 'rein'
	PAESE/O 'country'						
	QN/O 'somebody'				PROGETTO/O 'project'		PRESA/O 'grip'
	CORAGGIO/O 'courage'		LAVORO/O 'job'				
	NAVE/O 'ship'				SPERANZA/O 'hope'		TESTA/O 'head'
	CAMPO/O 'camp'						

The table illustrates the paradigmatic family of the verb *abbandonare* 'abandon' which consists of four distinct paradigms, one for each different sense of the verb as attested in the example base. For convenience, a single paradigm is represented as a two-column matrix. The left column contains the particular sense of *abbandonare* (identified by a number) that all patterns share. Such a common element is the "core" of the paradigm. The right column of the matrix consists of a list of "paradigmatic slots" which contain the elements that are left out of the core: namely, the different nouns which co-occur with a certain sense of abbandonare. For each noun, its syntactic role is indicated after a slash; e.g. "O" stands for 'object'. Nouns represent the disambiguating cues for each sense of *abbandonare* and are in parallel distribution with respect to its core (i.e. they are mutually substitutable in that context). Note that, for what concerns us here, this distributional equivalence is the only property they share, or, to put it differently, their parallel distribution relative to a given sense suggests that they represent a semantically coherent set of entities whose similarity remains implicit in the grouping.

Disambiguating cues, in their turn, give rise to further paradigms, when they combine with more than one verb sense, as illustrated in (2) for *figlio* 'son'.

(2)

FIGLIO 'son'	
paradigm E	
OBJECT	CRESCERE_$0_4 'grow'
	LEGITTIMARE_$0_1 'legitimate'
	PIANTARE_$0_3 'quit'
	RICONOSCERE_$0_3 'acknowledge'
	RIMPROVERARE_$0_1 'scold'
	RINNEGARE_$0_1 'disown'
	SCHIAVIZZARE_$0_1 'subjugate'

Unlike (1), (2) leaves its core unspecified as to its sense; conversley, sense specification characterises here each paradigmatic slot filler. Paradigmatic structures such as (1) and (2) are used by SENSE for making predictions about the most likely sense(s) of words in unknown sentences.

4.2 Paradigm extension

For word sense disambiguation, paradigm extension can informally be defined as follows: if a certain verb shares at least one paradigmatic slot with another verb, then – for lack of better evidence – the tentative assumption is made that the former verb inherits all paradigmatic slots of the latter; the same holds for noun paradigms.

To illustrate this, let us suppose that the unknown expression *abbandonare-figlio* 'abandon-son' is given to SENSE as a target expression. The system is told that *figlio* is the object of *abbandonare* and asked to disambiguate the sense of the verb. SENSE has to select any of the senses 0_1, 0_3, 0_4 or 0_5 of *abbandonare* attested in the example base. Consider first the hypothesis that *abbandonare* is used in sense 0_1. For SENSE to support this hypothesis paradigmatically, a verb-noun pattern has to be found where one of the fillers of the paradigmatic slots of *abbandonare_$0_1* (figure 1) co-occurs with one of the fillers of the paradigmatic slots of *figlio* as an object (figure 2). This case has been found as shown in figure (3):

(3)

ABBANDONARE$0_1	PAESE/O
	QUALCUNO/O
	.../O

RINNEGARE$0_1	**FAMIGLIA/O**
	FEDE/O
	.../O
	FIGLIO/O

where the paradigms *abbandonare_$0_1* and *rinnegare_$0_1* appear to share the paradigmatic slot *famiglia/O* (the shadowed box in 3). This means that the hypothesis *abbandonare_$0_1-figlio* is justified on the basis of paradigm extension. More procedurally, the target *abbandonare-figlio* is analogically mapped onto *abbandonare_$0_1-famiglia* on the basis of the distributional evidence that both *famiglia* and *figlio* can be the object of *rinnegare_$0_1* 'disown'. Note that the same interpretation is further supported by the intersection with two other paradigms, i.e *piantare_$0_1* 'quit' and *riconoscere_$0_3* 'recognize' both of which contain *qualcuno* 'somebody' as an object. Consider now *abbandonare* in the other senses. It appears that no paradigmatic evidence is found in the example base to support them: i.e. no paradigmatic slots are shared by these other senses of the verb *abbandonare* and the verb senses which co-occur with *figlio* as an object.

To sum up, paradigm extension for word sense disambiguation can be defined as follows: the disambiguating cue of a given word sense s_1 (e.g. *figlio* with respect to *rinnegare_$0_1*) is also used as a disambiguating cue of another word sense s2 (i.e. *abbandonare_$0_1*) if s_2 shares at least one disambiguating cue (e.g. *famiglia*) with s_1 (*rinnegare_$0_1*).

5. SENSE: first experimental results

Tests were carried out to assess the performance of a paradigm-based strategy in word sense disambiguation. The example base was automatically extracted (Montemagni 1995) from the Collins bilingual Italian-English dictionary in MRF (Picchi et al. 1992). We opted for the dictionary source rather than textual corpora since data acquired from the former offer the nonnegligible advantage of always referring to a specific sense of the headword (contrary to the other words in the acquired

pattern which are not specified for a particular sense). The example base, at the current stage, contains verb-noun pairs only, where the noun is either the subject or the object of the verb; clearly, the example base can be extended to contain other types of word co-occurrence pattern. Verb-noun pairs were all acquired from 3,353 verb entries, corresponding to 4,652 different senses. From this set of entries, 8,153 verb-noun co-occurrence patterns were extracted: 5,665 are verb-object patterns and 2,488 verb-subject ones.

The test corpus, i.e. the set of target expressions to be disambiguated, consists of 100 co-occurrence patterns extracted from unrestricted text; patterns already present in the example base are excluded. SENSE picks up the relevant sense in 67% of the cases, leaves the sense ambiguous in 24% of the cases, goes wrong with 9% of the target expressions. The overall accuracy rate is 88%.

The results are obtained by enforcing the following constraints: a) whenever more than one sense assignment is paradigmatically supported, the most widely supported assignment wins over the other one(s); b) the evidence supporting different sense assignments is weighted according to its generality/specificity, i.e. more specific evidence is given preference over semantically vaguer one; c) paradigm extension is made possible across paradigms with nouns playing different syntactic functions (either subject or object).

6. Final remarks

This paper shows how paradigm extension can be used for sense disambiguation in an example-based system. This mechanism, unlike classical sense disambiguation procedures, makes use of a structured network of words only some of which are semantically disambiguated. Thesaurical information is not a necessary prerequisite (although it cann be used conveniently) so that the system's ability to find relevant analogues is not forced into the straitjacket of a fixed conceptual hierarchy. In a preliminary evaluation of this strategy a high number of cases is successfully resolved with considerable accuracy. The sample of successful cases includes both literal and figurative usages, since conventional figurative usages, showing metonymical or metaphorical sense extensions, have been acquired from the dictionary source on a par with literal ones. The fact that SENSE can deal with both kinds of literal and figurative usages makes it robust enough to be able to cope with unrestricted text.

Notes

* For the specific concerns of the Italian Academy, Montemagni is responsible for sections 1, 2 and 3, Federici for sections 4, 4.1. and 6, Pirrelli for sections 4.2 and 5.

1 This is the approach to translation selection and word sense discrimination adopted by Michiels (1994).

References

Collins Giunti Marzocco, 1985, *English-Italian Italian-English Dictionary*, Collins Giunti Marzocco, London Firenze.

Dolan W., 1994, "Word sense ambiguation: Clustering related Senses", in Proceedings of COLING-94, Kyoto, Japan, pp. 712–716.

Federici S., Pirrelli V., 1994, "Linguistic Analogy as a Computable Process", in *Proceedings of NeMLaP*, Manchester, UK, pp. 8–14.

Garzanti, 1984, *Il Nuovo Dizionario Italiano Garzanti*, Garzanti, Milano.

Harley A., forthcoming, "Cambridge Language Survey: Semantic Tagger", in *Proceedings of the Pisa Acquilex Workshop, July 1994*.

Michiels A., 1994, An experiment in translation selection and word sense discrimination using the metalinguistic apparatus of two computerized dictionaries, Working Paper W-002, DECIDE MLAP Project 93/19.

Montemagni S., 1995, *Subject and Object in Italian Sentence Processing*, PhD Dissertation, Umist Manchester, UK.

Nagao M., 1992, "Some Rationales and Methodologies for Example-Based Approach", in *Proceedings of "International Workshop on Fundamental Research for the Future Generation of Natural Language Processing", 30–31 July 1992*, Manchester, UK, pp. 82–94..

Picchi E., Peters C., Marinai E., 1992, "The Pisa Lexicographic Workstation: the Bilingual Components", in *Proceedings of Euralex 92,* Tampere, Finland, pp. 277–285.

Pirrelli V., Federici S., 1994, "Derivational Paradigms in Morphonology", in *Proceedings of COLING-94*, Kyoto, Japan, pp. 234–242.