Computers and the Humanities **34:** 199–204, 2000.

ROMANSEVAL: Results for Italian by SENSE

STEFANO FEDERICI, SIMONETTA MONTEMAGNI and VITO PIRRELLI
Istituto di Linguistica Computazionale – CNR, Via della Faggiola 32, Pisa, Italy
(E-mail: {stefano,simo,vito}@ilc.pi.cnr.it)

Abstract. The paper describes SENSE, a word sense disambiguation system that makes use of different types of cues to infer the most likely sense of a word given its context. Architecture and functioning of the system are briefly illustrated. Results are given for the ROMANSEVAL Italian test corpus of verbs.

Key words: analogy-based NLP, semantic similarity, word sense disambiguation

1. Word Sense Disambiguation by SENSE

SENSE (Self-Expanding linguistic knowledge-base for Sense Elicitation) is a specialised version of a general purpose language-learning system (Federici and Pirrelli, 1994; Federici et al., 1996) tailored for sense disambiguation (Federici et al., 1997, 1999).

SENSE belongs to the family of example-based Word Sense Disambiguation (WSD) systems as it assigns, to an ambiguous word token W_k in a target context C_j, the sense with which W_k is tagged in another context similar or identical to C_j. Hereafter, a target word token W_k and its context C_j will jointly be referred to for convenience as the "input pattern". Knowledge of the way the senses of W_k appear in context comes from a repertoire of examples of use of word senses, or "Example Base" (*EB*). The *EB* of SENSE contains three basic types of such contexts: (i) subcategorisation patterns (e.g. an infinitival construction governed by a given sense of W_k), (ii) functionally annotated word co-occurrence patterns (e.g. the typical objects of W_k if the latter is a verb), (iii) fixed phraseological expressions.

The similarity of a target context C_j to the contexts in *EB* is measured differently depending on i, ii or iii. Contexts of type i and iii are dealt with through simple pattern-matching: C_j is either identical to another context in *EB* where W_k occurs (in which case the sense of W_k in that context is selected), or no answer is given. On the other hand, when C_j is part of a functionally annotated word co-occurrence pattern (type ii above) then similarity does not necessarily require full identity. This means that when SENSE fails to find an *EB* context identical to C_j, it tries to match C_j against a semantically similar context. Semantic similarity is assessed through a “proportionally-based” similarity measure briefly illustrated in section 1.1.

SENSE outputs either one sense of W_k in C_j (if only one sense is supported by *EB*), or a ranked list of possible alternative senses. The ranking procedure is sketched in section 1.2.

1.1. PROPORTIONALLY-BASED SEMANTIC SIMILARITY

The key formal notion used by SENSE to compute similarity between non identical (functionally annotated) contexts is “proportional analogy”. To illustrate, suppose that SENSE has to guess the sense of the Italian verb *accendere* in the pair *accendere-pipa*/O ‘light–pipe’ (where ‘pipe’ is tagged as a direct object) and that the input pattern in question is not already present in *EB*. Then the system goes into *EB* looking for functionally annotated patterns entering a proportion such as the following:

t_1	:	t_2	=	t_3	:	t_4
fumare$_1$		*fumare*$_1$		*accendere*$_1$		*accendere*$_?$
sigaretta$_1$/O		*pipa*$_1$/O		*sigaretta*$_1$/O		*pipa*$_1$/O
‘smoke–cigarette/O’		‘smoke–pipe/O’		‘light–cigarette/O’		‘light–pipe/O’

The proportion involves three *EB* verb-object pairs where the verb is sense-tagged (t_1, t_2 and t_3), plus the input pattern *accendere-pipa*/O (t_4 or “target term”). The proportion is solved by assigning *accendere* in t_4 the sense *accendere*$_1$, by analogical transfer from t_3 (or “transfer term”). Intuitively, the proportion suggests that the sense of *accendere* in the input pattern is likely to be the same as the one in the pattern *accendere*$_1$-*sigaretta*$_1$, since *pipa*$_1$ and *sigaretta*$_1$ are found to be in complementary distribution relative to the same sense *fumare*$_1$ of the verb *fumare* ‘smoke’. t_1, or “pivot term”, plays the role of linking the target with the transfer term. We can say that analogical proportions are able to transfer word senses across sense-preserving contexts. Note further that here the similarity between contexts depends on W_k (e.g., *accendere* in the case at hand): ‘pipe’ and ‘cigarette’ are semantic neighbours only relative to some verbs (e.g. ‘smoke’ or ‘light’, as opposed to, e.g., ‘roll’ or ‘fill’).[1]

Observe that, in the analogical proportion above, nouns stand in the same syntactic relation to verbs. In other cases, however, clusters of nouns which function, say, as the object of a given verb sense also function as typical subjects of other related verb senses. This is captured through proportions of verb-noun pairs involving syntactically-asymmetric constructions, as exemplified below:

t_1	:	t_2	=	t_3	:	t_4
rappresentare$_1$		*rappresentare*$_1$		*attaccare*$_1$		*attaccare*$_?$
quadro$_1$/S		*foto*$_1$/S		*quadro*$_1$/O		*foto*$_1$/O
‘show painting/S’		‘show photo/S’		‘hang_up painting/O’		‘hang_up photo/O’

In the proportion, $foto_1$ 'photo' and $quadro_1$ 'painting' are semantically similar due to their both being subjects of the same sense of the verb *rappresentare* 'represent' ($rappresentare_1$). This similarity is supposed to proportionally carry over to the case of the same two nouns being used as typical objects of *attaccare* 'hang'. The inference is made that the sense of *attaccare* in the target term is $attaccare_1$, by analogy to the transfer term $attaccare_1$-$quadro_1$/O.

When proportions are found which support more than one sense interpretation of W_k, alternative interpretations are weighted according to their analogy-based support. The weight reflects: (i) number of proportions supporting a given sense interpretation and (ii) semantic entropy of the words in the pivot terms of the supporting proportions (calculated according to the Melamed (1997) definition, i.e. as $\log_2(\text{freq}(W_k))$ where "freq" counts the number of different functionally annotated *EB* patterns containing W_k).[2]

1.2. MULTI-CUE WSD AND RANKING OF RESULTS

We deal here with the way SENSE weighs multiple sense assignments depending on what type of *EB* context supports them. Input patterns are projected onto *EB* by looking for matching phraseological contexts first (if any), and then for functionally annotated word co-occurrence patterns. Syntactic frames are looked for only as a last resort.

Existence of an *EB* lexical pattern (type ii or iii in section 1) identical to the input pattern is always given full credit, and the corresponding W_k sense is selected. For lack of identical lexical evidence, similar contexts are searched for through analogical proportions. If more than one sense is proportionally supported, the one with the heaviest analogical weight (section 1.1) is selected. Subcategorisation patterns are resorted to only when lexical evidence is inconclusive.

2. Experimental Setting

In the experiment reported here, SENSE is asked to assign senses to verb occurrences in the ROMANSEVAL test corpus on the basis of a bi-partitioned *EB*.

2.1. THE TEST CORPUS

The ROMANSEVAL test corpus contains 857 input patterns of 20 different polysemous verbs. The verbs show different degrees of polysemy: the number of senses ranges from the 16 senses of *passare* 'pass' to the 2 senses of *prevedere* 'foresee'; on average, each verb has 5 different senses. Input patterns are fed into SENSE after a parsing stage (see Federici et al., 1998a,b) which outputs them as syntactically annotated patterns. These patterns are compatible with any of the three types of context in *EB* (section 1).

2.2. THE EXAMPLE BASE

In this experiment, SENSE uses a bi-partitioned *EB*. The first partition is a generic resource containing 17,359 functionally annotated verb-noun patterns (6,201 with subject, and 11,148 with object), with no indication of sense for either member of the pair. We will hereafter refer to this partition as the "unsupervised tank". These patterns were automatically extracted (Montemagni, 1995) from both definitions and example sentences of the verb entries of a bilingual Italian-English dictionary (Collins, 1985) and a monolingual Italian dictionary (Garzanti, 1984). They represent the typical usage of 3,858 different verbs, each exemplified through a comparatively sparse number of patterns (on average 4.5 per verb). Although these patterns were originally sense-tagged on the verb, we could not use these tags, since (a) they referred to sense distinctions coming from different dictionaries, and (b) they could not easily be mapped onto ROMANSEVAL sense distinctions.

The second partition is specific to each test word W_k: it contains a number of patterns attesting the different senses of W_k as defined by ROMANSEVAL. The patterns include: (i) patterns originally belonging to the unsupervised tank and manually sense-tagged; (ii) patterns extracted from the lexicon adopted in ROMANSEVAL as a reference resource. This partition contains a comparatively small number of patterns (an average of 31.6 per W_k) exemplifying an average of 6 contexts of use of each of W_k senses. Typical word co-occurrence patterns form 87% of the partition, subcategorisation patterns 10% and phraseological expressions about 3%. Note that only W_k is sense-tagged in these patterns which thus act as "sense seeds" of W_k (Yarowsky, 1995).

2.3. ANALOGICAL PROPORTIONS WITH A BI-PARTITIONED EB

In this section we briefly illustrate the way the bi-partite *EB* described above is used to establish analogical proportions. Given an input pattern, SENSE tries to establish analogical proportions by looking for the transfer term in the partition of sense seeds, while t_1 and t_2 are looked for in the unsupervised tank. Proportions of this sort are intuitively less constrained than those illustrated in section 1.1, since nouns in the proportion are no longer proved to be in complementary distribution relative to the same verb sense, but simply relative to the same verb. Relaxing this constraint was necessary since, as pointed out above, our *EB* combines sense distinctions coming from different dictionaries. This evaluation protocol amounts to testing analogy-based WSD in a fully unsupervised way.

3. Results

Results of the experiment are encouraging. Recall, calculated as the number of correct answers relative to the total number of input patterns, is 67% and precision 85%. Correct answers include: (a) one-sense answers (over 95% of the total); (b) more-than-one-sense answers, when the correct sense is given the topmost weight together with a subset of the attested senses of W_k in *EB*. SENSE fails on 11% of the input patterns. Input patterns for which SENSE yields no answer amount to 22% of the total. Almost half of them (i.e. 86 out of 192) contain context words missing in *EB* for which no proportion could possibly be established.

It is interesting to consider the individual contribution of each context type (see section 1) to the disambiguation task: 72% of SENSE correct answers are based on lexico-semantic patterns (either fixed phraseological expressions or typical word co-occurrence patterns representative of the selectional preferences of a specific verb sense); 28% are based on subcategorisation information. Analogical proportions contribute 52% of correct sense assignments.[3] Note finally that, in the test sample, more-than-one-sense answers are always due to subcategorisation patterns.

4. Concluding Remarks

In this paper, we illustrated an analogy-based system for WSD capable of dealing with different types of linguistic evidence (syntactic and lexico-semantic), and report the results obtained on the ROMANSEVAL test bed. One of the most innovative features of the system is that similarity between contexts is computed through analogical proportions which, in the reported experiment, are minimally constrained, i.e. they are based on a handful of sense-tagged contexts (or sense seeds) reliably extended through a set of untagged data (forming an unsupervised tank). This amounts to testing analogy-based WSD in a fully unsupervised mode, and it has an obvious bearing on the scalability and exportability of the proposed method. For a given W_k, one can "plug", into *EB*, different sense subdivisions (e.g. exhibiting varying degrees of granularity), and disambiguate W_k in context accordingly. Moreover, the unsupervised tank can either be extended through new lexical patterns extracted from unrestricted texts, or specialised through addition of domain-specific contexts.

Notes

[1] For a comparison between this operational notion of similarity and alternative proposals used in other analogy-based systems for WSD, the reader is referred to Federici et al. (1999).

[2] A detailed discussion of the weighting procedure can be found in Federici et al. (1997, 1999).

[3] This figure is obtained by forcing SENSE to disambiguate all input patterns proportionally, i.e. pretending that no input pattern was already present in the partition of sense seeds.

References

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

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

Federici, S., S. Montemagni and V. Pirrelli. "Analogy and Relevance: Paradigmatic Networks as Filtering Devices". In *Proceedings of NeMLaP*. Ankara, Turkey, 1996, pp. 13–24.

Federici, S., S. Montemagni and V. Pirrelli. “Inferring semantic similarity from Distributional Evidence: an Analogy-based Approach to Word Sense Disambiguation”. In *Proceedings of the ACL/EACL Workshop “Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications”*. Madrid, Spain, 12 July 1997.

Federici, S., S. Montemagni and V. Pirrelli. “Chunking Italian: Linguistic and Task-oriented Evaluation”. In *Proceedings of the LRE-98 Workshop on Evaluation of Parsing Systems*. Granada, Spain, 26 May 1998, 1998a.

Federici, S., S. Montemagni, V. Pirrelli and N. Calzolari. “Analogy-based Extraction of Lexical Knowledge from Corpora: the SPARKLE Experience”. In *Proceedings of LRE-98*, Granada, Spain, 28–30 May 1998, 1998b.

Federici, S., S. Montemagni and V. Pirrelli. “SENSE: an Analogy-based Word Sense Disambiguation System”. In *Natural Language Engineering*, 1999.

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

Melamed, D. “Measuring Semantic Entropy”. In *Proceedings of SIGLEX Workshop on Tagging Text with Lexical Semantics: why, what and how?*, ANLP’97, Washington, USA, 4–5 April 1997.

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

Yarowsky, D. “Unsupervised word sense disambiguation rivaling supervised methods”. In *Proceedings of ACL ’95*, Cambridge, MA, June 1995, pp. 189–196.