Experiments on the Validation of Sense Annotations Assisted by Lexical Chains

Roberto Navigli
Dipartimento di Informatica
Università di Roma "La Sapienza"
Roma, Italy
navigli@di.uniroma1.it

Abstract

It is widely recognized that the annotation of texts with senses from a computational lexicon is a complex and often subjective task. We propose the use of lexical chains, specifically semantic interconnections, to support validators in the difficult task of assessing the quality of sense assignments. We provide a two-fold experimental evaluation of our approach applied to the validation of manual annotations from the SemCor corpus, and we further assess the method on automatic annotations from the English all-words Senseval 3 competition.

Introduction

Sense annotation is the task of assigning senses chosen from a computational lexicon to words in context. This is a task where both machines and humans find it difficult to reach an agreement. The problem depends on a variety of factors, ranging from the inherent subjectivity of the task to the granularity of sense discretization, coverage of the reference dictionary, etc.

The problem of validation is even amplified when sense tags are collected through acquisition interfaces like the *Open Mind Word Expert* (Chklovski and Mihalcea, 2002), due to the unknown source of the contributions of possibly unskilled volunteers.

Strategies like *voting* for automatic sense annotations and the use of *inter-annotator agreement* with adjudication for human sense assignments only partially solve the issue of disagreement. Especially when there is no clear preference towards a certain word sense, the final choice made by a judge can be subjective, if not arbitrary. This is a case where analysing the intrinsic structure of the reference lexicon is essential for producing a consistent decision. A lexicographer is indeed expected to review a number of related dictionary entries in order to adjudicate a sense coherently. This work can be tedious, time-consuming, and often incomplete, due to the complex structure of the resource. As a result, inconsistent choices can be made.

In this paper, we present and evaluate a knowledge-based method for assisting the validation of both manual and automatic sense annotations. The paper is organized as follows: first, we introduce lexical chains and semantic interconnections (Section 1), we illustrate our method for the validation of sense annotations (Section 2), and we evaluate the approach applied to both manual and automatic annotations (Section 3). Finally, in Section 4 we present some conclusions.

1 Lexical Chains and Semantic Interconnections

Semantic networks are a graphical notation developed to represent knowledge explicitly as a set of conceptual entities and their interrelationships. The availability of wide-coverage computational lexicons like WordNet (Fellbaum, 1998), as well as semantically annotated corpora like SemCor (Miller et al., 1993), has certainly contributed to the exploration and exploitation of semantic graphs for several tasks like the analysis of the lexical text cohesion (Morris and Hirst, 1991), word sense disambiguation (Agirre and Rigau, 1996; Mihalcea and Moldovan, 2001), ontology learning, etc.

Lexical chains (Morris and Hirst, 1991), inspired by the notion of cohesion in discourse, are

sequences of words $w_1, \ldots, w_n$ in a text that represent the same topic, i.e. such that w_i is related to w_{i+1} by a lexico-semantic relation (e.g. hypernymy, meronymy, etc.). Subsequent works (e.g. Mihalcea and Moldovan (2001)) further develop this idea by providing knowledge-based approaches to Word Sense Disambiguation. Given a word context $\sigma = w_1, w_2, \ldots, w_n$ and a lexical knowledge base, these approaches tend to select those configurations of senses $\hat{s}_{w_1}, \hat{s}_{w_2}, \ldots, \hat{s}_{w_n}$ that maximize the degree of mutual interconnection, according to a measure of connectivity, that is $\hat{s}_w = \arg\max_{s_w \in Senses(w)} f(s_w, \sigma)$, where f is a function of the lexical chains connecting s_w to the senses chosen for σ.

Recently, a knowledge-based algorithm for Word Sense Disambiguation, called *Structural Semantic Interconnections*[1] (SSI) (Navigli and Velardi, 2005), has been shown to provide interesting insights into the choice of word senses by providing structural justifications in terms of a specific kind of lexical chains, called semantic interconnections.

A *semantic interconnection pattern* is a relevant sequence of edges selected according to a context-free grammar, i.e. a path connecting a pair of word senses (dark nodes in Figure 1), possibly including a number of intermediate concepts (light nodes in Figure 1). The SSI algorithm exploits a lexical knowledge base, based on the WordNet lexicon and enriched with a number of *relatedness* relations, connecting pairs of related word senses. The enrichment is based on the acquisition of collocations from existing resources (like the Oxford Collocations, the Longman Language Activator, collocation web sites, etc.). Each collocation is mapped to the WordNet sense inventory in a semi-automatic manner (Navigli, 2005) and transformed into a *relatedness* edge.

We choose the connectivity function f as the normalized sum of the inverse length of interconnections (i.e. the contribution of a single connection $s_w \rightarrow^* s_{w'}$ is $\frac{1}{length(s_w \rightarrow^* s_{w'})}$) between s_w and the other senses chosen in context. Given the sense configuration $\hat{s}_{w_1}, \hat{s}_{w_2}, \ldots, \hat{s}_{w_n}$ that maximizes the degree of mutual interconnection, word $w \in \sigma$ is assigned the word sense $\hat{s}_w$ if the normalized sum of the contributions coming from the other senses $\hat{s}_{w'}$ ($w' \in \sigma$, $w' \neq w$) is over a fixed threshold.

For example, if the context of words to be disambiguated is [*cross-v*, *street-n*, *intersection-n*], the senses chosen by SSI with respect to WordNet are: [*cross-v#1*, *street#2*, *intersection#2*][2], supported – among the others – by the pattern *intersection#2* $\xrightarrow{part-of}$ *road#1* $\xleftarrow{kind-of}$ *thoroughfare#1* $\xleftarrow{kind-of}$ *street#2*. An excerpt of the manually written context-free grammar encoding valid semantic interconnection patterns for the WordNet lexicon is reported in Table 1. The grammar allows to avoid the recognition of unwanted patterns causing a deep shift of meaning (e.g. *universe#1* $\xrightarrow{kind-of}$ *natural object#1* $\xrightarrow{kind-of}$ *object#1* $\xrightarrow{has-kind}$ *commodity#1* $\xrightarrow{has-kind}$ *merchandise#1*, or *job#1* $\xrightarrow{related-to}$ *money#1* $\xrightarrow{related-to}$ *coin#1* $\xrightarrow{related-to}$ *metal#1*). For further details the reader can refer to the literature (e.g. Navigli and Velardi (2005)).

Table 1: An excerpt of the context-free grammar for the recognition of semantic interconnections.

$S \rightarrow S'S_1 \vert S'S_2 \vert S'S_3$
(start rule)
$S' \rightarrow e_{nominalization} \vert e_{pertainymy} \vert \epsilon$
(part-of-speech jump)
$S_1 \rightarrow e_{kind-of}S_1 \vert e_{part-of}S_1 \vert e_{kind-of} \vert e_{part-of}$
(hyperonymy/meronymy)
$S_2 \rightarrow e_{kind-of}S_2 \vert e_{relatedness}S_2 \vert e_{kind-of} \vert e_{relatedness}$
(hypernymy/relatedness)
$S_3 \rightarrow e_{similarity}S_3 \vert e_{antonymy}S_3 \vert e_{similarity} \vert e_{antonymy}$
(adjectives)

In this paper, we aim at showing that knowledge-based, untrained WSD algorithms founded on the concept of lexical chains, and specifically on semantic interconnections, can help speed up the task of validating sense annotations. As illustrated in the following, semantic interconnections are an important requirement for this purpose in that the outcome of the algorithm applied to a sentence σ can be visualized in terms of semantic graphs representing the patterns connecting the suggested senses.

[1] SSI is an online WSD algorithm available at http://lcl.di.uniroma1.it/ssi.

[2] We indicate a word sense with the convention *w-p#i*, where w is a word, p its part of speech (n for nouns, a for adjectives, v for verbs, r for adverbs) and i its sense number in the WordNet inventory. For readability, in the following we omit the noun part of speech.

2 Supporting Validation with Semantic Interconnection Patterns

The task of validating sense annotations can be defined as follows: let w be a word in a sentence σ, previously tagged by a set of annotators $A = \{a_1, a_2, ..., a_n\}$ each providing a sense for w, and let $S = \{s_1, s_2, ..., s_m\} \subseteq Senses(w)$ be the set of senses chosen for w by the annotators in A, where $Senses(w)$ is the subset of senses of w in the reference inventory (we adopt WordNet). A validator is asked to validate, that is to adjudicate a sense $s \in Senses(w)$ for a word w over the others. Notice that the annotators in A can be either human or automatic, depending upon the purpose of the exercise.

Given a set of words with disagreement $W \subseteq \sigma$, we apply SSI to W by taking into account for disambiguation only the senses in S (i.e. the set of senses selected by the annotators), and using as a fixed context the agreed senses chosen for the words in $\sigma \setminus W$.

In the following subsections, we describe the application of our method to the validation of manual and automatic annotations, and we discuss cases of uncertain applicability.

2.1 Validating Manual Annotations

Consider the following sentence:

(a) We crossed the street near the intersection

All the occurrences of the phrase *cross the street* in the SemCor corpus are tagged with the first sense of *street* (defined in WordNet as *a thoroughfare (usually including sidewalks) that is lined with buildings*), but it is clear, from the definition of the second sense (*the part of a thoroughfare between the sidewalks; the part of the thoroughfare on which vehicles travel; "be careful crossing the street"*), that a pedestrian crosses that part of the thoroughfare between the sidewalks. Though questionable, this is a subtlety made explicit in the dictionary and reinforced by the usage example of sense #2 above.

Suppose two annotators agreed on the senses of *cross* and *intersection*, but disagreed on the word *street*, choosing respectively the first and the second sense from the WordNet inventory.

The application of the SSI algorithm to sentence (a) leads to the suggestion of the second sense as a solution to this disagreement. This suggestion is supported by a number of semantic interconnections according to the grammar in Table 1. Figure 1(a) shows some interconnections suggested by the algorithm.

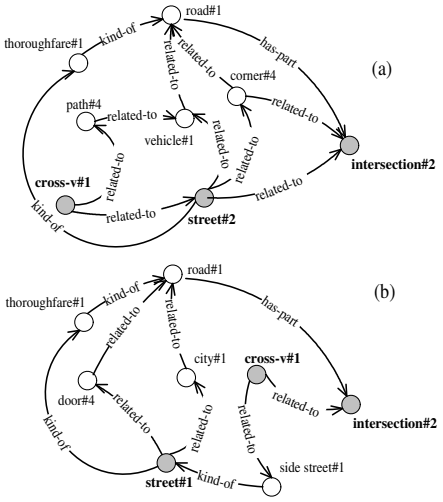

Figure 1: (a) Semantic interconnection patterns supporting the choice of sense #2 of *street* in sentence (a). (b) Some interconnections supporting the choice of *street#1* for the same sentence.

Semantic interconnections reflect the fine granularity of the inventory, as they are expressions of the lexical knowledge base from which they are extracted. In fact, the choice of $street\#1$ still produces good semantic interconnections, as illustrated in Figure 1(b), but the overall ranking of this sense selection, i.e. the degree of overall connectivity of the resulting graph, is smaller than that obtained for $street\#2$.

As a second example, consider the WordNet definition of *motorcycle*:

(b) Motorcycle: a motor vehicle with two wheels and a strong frame

In the Gloss Word Sense Disambiguation task at Senseval-3 (Litkowski, 2004), the human annotators assigned the first sense to the word *frame* (*a structure supporting or containing something*), unintentionally neglecting that the dictionary encodes a specific sense of *frame* concerning the structure of objects (e.g. vehicles, buildings, etc.). In fact, according to WordNet, a *chassis#3* is a kind of *frame#6* (*the internal supporting structure that gives an artifact its shape*), and is also part of a *motor vehicle#1*. While regular polysemy holds between sense #1 and #6, there is no justification

for the former choice, as it does not refer to vehicles at all (as reflected by the lack of semantic interconnection patterns concerning *frame#1*).

From these two real-world cases, it is evident that semantic interconnections can point at inconsistent, though acceptable, choices made by human annotators due, among others, to the fine granularity of the sense inventory and to regular polysemy.

Apart from tagging mistakes, most of the cases of disagreement between manual annotators is due to the fine granularity of the lexicon inventory. We recognize that subtle distinctions, like those encoded in WordNet, are rarely useful in any NLP application, but as a matter of fact WordNet is at the moment the *de facto* standard within the research community, as no other computational lexicon of that size and complexity is freely available.

2.2 Validating Automatic Annotations

While the task of manual annotation is mostly restricted to lexicographers, the automatic annotation of texts (especially, web pages) is gaining a growing popularity in the Semantic Web vision (Berners-Lee, 1999). In order to perform automatic tagging, one or more word sense disambiguation systems are applied, resulting in a semantically-enhanced resource. Unfortunately, even when dealing with restricted sense inventories or selected domains, automated systems can make mistakes in the sense assignment, also due to the difficulty in training a supervised program with a sufficient number of annotated instances and again due to the fine granularity of the dictionary inventory.

There are also cases in which an automatic disambiguator chooses a partially justifiable, but incorrect interpretation for words in context. Consider for instance the sentence from the Senseval-3 English all-words competition:

(c) The *driver* stopped swearing at them, *turned* on his *heel* and went back to his *truck*

A partial interpretation of *driver* and *heel* can be provided in the golf domain (a *heel#6* is part of a *driver#5*). This can be a reasonable choice for a word sense disambiguator, but the overall semantic graph exposes a poor structural quality. A different choice of senses pointed out by stronger semantic interconnections (*driver* as an operator of a vehicle and *heel* as the back part of the foot) provides a more interconnected structure (among others, $driver\#1 \xrightarrow{related-to} motor\ vehicle\#1 \xleftarrow{kind-of} truck\#1$, *turn-v*$\#1 \xrightarrow{related-to} heel\#2$, etc.).

2.3 Weaknesses of the approach

It can happen that semantic interconnection patterns convey weak suggestions due to the lack of structure in the lexical knowledge base used to extract patterns like those in Table 1. In that case, the validator is expected to reject the possible suggestion and make a more reasonable choice. As a result, if no interesting suggestion is provided to the validator, it is less likely that the final choice will be inconsistent with the lexicon structure. A typical example is:

(d) A *payment* was made last week.

WordNet encodes two senses of *payment*: the sum of money paid (sense #1) and the act of paying money (sense #2). Such regular polysemy makes it hard to converge on a sense choice for *payment* in sentence (d). This difficulty is also manifested in the annotations of similar expressions involving *make* and *payment* within SemCor. Furthermore, apart from the distinction between the act of doing the action and the amount of money paid, there are not many structural suggestions allowing to distinguish between the two senses. Semantic interconnections cannot help the validator here, but any choice will not violate the structural consistency of the lexicon.

3 Evaluation

The objective of this section is to show that semantic interconnections constitute a good support for a validator in the detection of bad or inconsistent annotations. We assessed the method for both manual (Section 3.1) and automatic annotations (Section 3.2). In Section 3.3 we discuss the experiments. The evaluations are all based on standard test sets.

3.1 Evaluating the Validation of Manual Annotations

We made two experiments for assessing the suggestions provided by SSI for validating manual annotations, both based on the semantically-tagged SemCor corpus (Miller et al., 1993).

In a first experiment, 1,000 sentences were uniformly selected from the set of documents in the SemCor corpus. For each sentence

$\sigma = w_1 w_2 \ldots w_n$ annotated in SemCor with the senses $s_{w_1} s_{w_2} \ldots s_{w_n}$ ($s_{w_i} \in Senses(w_i), i \in \{1, 2, \ldots, n\}$), we randomly identified a word $w_i \in \sigma$, and chose at random a different sense $\overline{s}_{w_i}$ for that word, that is $\overline{s}_{w_i} \in Senses(w_i) \setminus \{s_{w_i}\}$. In other words, we simulated *in vitro* a situation in which an annotator provides an appropriate sense and the other selects a different sense. The random factor guarantees the uniform distribution over the test set of all the possible degrees of disagreement between sense annotators (ranging from regular polysemy to homonymy).

We applied SSI to the annotated sentences and evaluated the performance of the approach in suggesting the appropriate choice for the words with disagreement. We assessed both *precision* (the number of correct suggestions over the overall number of suggestions from SSI) and *recall* (the number of correct suggestions over the total number of words to be validated). The results are reported in Table 2 for nouns, adjectives, and verbs (we neglected adverbs as very few interconnections can be found for them).

Table 2: Results on a test set of 1,000 sentences from SemCor (one disagreed word per sentence chosen at random).

	Precision	**Recall**
Nouns	75.80% (329/434)	63.75% (329/516)
Adjectives	74.19% (46/62)	22.33% (46/206)
Verbs	65.64% (107/163)	43.14% (107/248)
Total	73.14% (482/659)	49.69% (482/970)
Baseline	50.00%	50.00%

The experiment shows that evidences of inconsistency due to inappropriate annotations are provided with good precision (we fix the baseline as the chance, that is we have 50% of probability to provide the appropriate sense for each word). The overall F1 measure (calculated as $\frac{2 \cdot p \cdot r}{p+r}$) is 59.18%. The improvement in precision over the baseline is statistically significant ($p < 0.01$).

Notice that this test bed differs from the typical evaluation of Word Sense Disambiguation tasks, like the Senseval exercises[3], in that we are assessing words w whose sense inventory is restricted to the set of senses $\{s_w, \overline{s}_w\}$.

The low recall obtained for verbs, but especially for adjectives, is due to a lack of connectivity in the lexical knowledge base, when dealing with connections across different parts of speech.

[3] http://www.senseval.org

Therefore, we repeated the experiment with a number of variations in the initial test set, by simulating for each sentence a disagreement on:

(α) the two most ambiguous words;

(β) the two least ambiguous words;

(γ) two words chosen at random;

(δ) three words chosen at random.

The results, shown in Table 3, provide interesting insights. First, validating sense annotations of highly polysemous words is more advantageous. In fact, in the case (α) the two senses selected for each word convey meanings which are more likely to be distant than in low-polysemy words (case (β)). As expected, the random choice strategy (γ) provides intermediate results between (α) and (β).

The second interesting remark is that when applying SSI to sentences with a number of disagreements per sentence greater than 1, the precision slightly decreases, and the recall increases, but both measures do not vary significantly (compare the figures in Table 2 with the results for (γ) and (δ) in Table 3).

Finally, we studied how the sentence size affects the results of case (γ). Figure 2 reports the results (similar figures can be obtained for the other cases). The figure highlights better, stable performances for a sentence size ranging between 8 and 12, which includes most of the sentences in our test set (754 sentences out of 1,000).

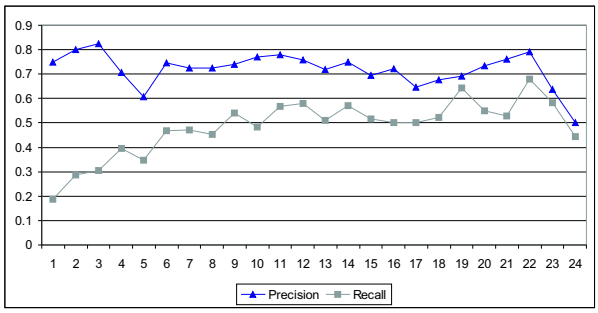

Figure 2: A graph of the performance by sentence size for case (γ).

In a second experiment, we applied semantic interconnections *in vivo* to a set of sentences that were used as a test set in the context of the

Table 3: Figures on variations of the first experiment (1,000 sentences).

	(α)				(β)			
	Precision		**Recall**		**Precision**		**Recall**	
Nouns	81.04%	(637/786)	67.40%	(637/945)	73.34%	(644/878)	60.98%	(644/1056)
Adjectives	93.33%	(126/135)	39.87%	(126/316)	66.94%	(81/121)	17.65%	(81/459)
Verbs	67.82%	(333/491)	47.77%	(333/697)	67.59%	(146/216)	46.49%	(146/314)
Total	77.62%	(1096/1412)	55.97%	(1096/1958)	71.68%	(871/1215)	47.62%	(871/1829)
	(γ)				(δ)			
	Precision		**Recall**		**Precision**		**Recall**	
Nouns	73.88%	(645/873)	62.25%	(645/1036)	72.72%	(925/1272)	62.03%	(925/1491)
Adjectives	81.10%	(103/127)	26.27%	(103/392)	77.14%	(135/175)	24.15%	(135/559)
Verbs	65.85%	(216/328)	45.56%	(216/474)	70.49%	(344/488)	50.36%	(344/683)
Total	72.59%	(964/1328)	50.68%	(964/1902)	72.55%	(1404/1935)	51.37%	(1404/2733)

MultiSemCor project[4], an English/Italian parallel version of the SemCor corpus (Bentivogli et al., 2004). In producing the test set, the lexicographers discovered a number of errors in the original SemCor annotations. Exploiting such divergences, we assessed the quality of the validation suggestions provided by SSI for each word with disagreement between the original SemCor annotations and those provided by the lexicographers working on MultiSemCor, assuming that the latters are the appropriate ones.

The original test set consisted of 119 words, 14 of which excluded because of a reconciliation of the disagreements due to the conversion of the data from WordNet 1.6 and 2.0. The test set consisted therefore of 105 words, contained in a total of 82 sentences from 4 SemCor annotated texts. We report the results in Table 4 (the test set includes two adverbs).

Table 4: Results on 82 SemCor sentences with 105 annotation errors.

	Precision	**Recall**
Nouns	64.00% (32/50)	58.18% (32/55)
Adjectives	40.00% (2/5)	11.76% (2/17)
Verbs	59.26% (16/27)	53.33% (16/30)
Total	61.45% (51/83)	49.04% (51/104)
Baseline	50.00%	50.00%

These figures are worse than those of the first experiment. This is due to the fact that in this case semantic interconnections have a minor impact, because the distinctions between the annotations from the original SemCor and those of MultiSemCor are often very subtle. Furthermore, the sample size is too small to provide a meaningful assessment. Our approach is still useful in a case like this, where a manual, visual inspection of the semantic interconnections can help the validator in accepting or discarding the suggestions provided[5], thus guaranteeing consistency with respect to the reference lexicon.

3.2 Evaluating the Validation of Automatic Annotations

For assessing semantic interconnections applied to the validation of automatic annotations, we chose the Senseval-3 corpus for the English all-words task (Snyder and Palmer, 2004). The task required WSD systems to provide a sense choice for a total of 2,081 content words in a set of 301 sentences from the fiction, news story, and editorial domains.

For our experiments, we focused on the outcome of the three best-ranking systems – GAMBL (Decadt et al., 2004), SenseLearner (Mihalcea and Faruque, 2004), and Koc University (Yuret, 2004) – and selected the subset of the sentences including one or more words with disagreement between the systems and such that at least one system made the appropriate sense choice according to the manual tagging provided by the organizers. We excluded from our test set any word included in our extended stopwords (e.g., *such*, *something*, etc.), resulting in a final figure of 411 disagreed words in 197 sentences.

The application of SSI to this test set led to the figures in Table 5 (the overall F1 measure was 52.51%). In the table, we compare our results with the chance baseline (calculated by taking into account the number of distinct answers given for each sentence by the three systems), the first sense heuristic (i.e. the choice of the most frequent sense in SemCor), and the best-performing Senseval system. The precision improvement over

[4]MultiSemCor is available online at: http://multisemcor.itc.it.

[5]As discussed in Section 4, our approach has been implemented as an online, freely available application.

the baseline is statistically significant ($p < 0.01$). Although the improvements of 6.5% and 5.04%, respectively, over the first sense and the best Senseval system are not statistically significant, we remark that we are comparing our method with the best supervised approaches. Furthermore, the (even smaller) difference in performance between the first sense and the best senseval system is not statistically significant as well.

Table 5: Results on 197 sentences from the Senseval-3 all words task (411 disagreed words).

	Precision	**Recall**
Nouns	54.59% $(113/207)$	51.13% $(113/221)$
Adjectives	55.88% $(19/34)$	38.00% $(19/50)$
Verbs	63.21% $(67/106)$	47.86% $(67/140)$
Total	57.35% $(199/347)$	48.42% $(199/411)$
Chance	47.28%	47.28%
First sense	$50.85\%(209/411)$	$50.85\%(209/411)$
Best senseval	$52.31\%(215/411)$	$52.31\%(215/411)$

An additional parameter to evaluate in this second experiment was the number of disagreements per sentence. The distribution of the 411 disagreed words over the sentences is illustrated in Figure 3.

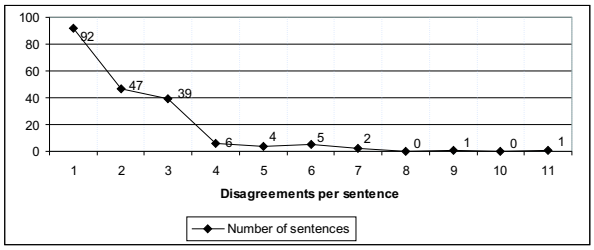

Figure 3: Distribution of the number of disagreements in the subset of the Senseval-3 test set.

We noticed that several small-size sentences in the test set are very difficult to be disambiguated even for humans. Examples of such sentences are (content words in italic): that*'s* what the *man said*; *was* the *man drunk* or *crazy* or both?; I*'m* just *numb*, etc. This observation led us to study the correctness of the suggestions provided by structural semantic interconnections for sentences whose size was over a fixed threshold. The results are shown in Table 6, when the threshold for the sentence size $|\sigma|$ ranges between 3 and 7.

3.3 Discussion

Many lexicographic studies established that human annotators cannot distinguish well between too fine-grained senses (e.g. Edmonds and Kilgariff (2002)). The ceiling of about 80% inter-annotator agreement (at least for English) was confirmed also in preparing the latest Senseval exercises. Assuming an average sentence size $|\sigma| = 10$ (a figure consistent with our experiments), we can reasonably suppose an average disagreement on two words in σ, a case similar to (γ), where our approach largely beats the baseline.

Table 6: Results thresholded on the sentence size.

$\lvert\sigma\rvert$	**Precision**	**Recall**
≥ 3	57.35% $(199/347)$	49.14% $(199/405)$
≥ 4	57.74% $(194/336)$	50.65% $(194/383)$
≥ 5	58.01% $(181/312)$	52.01% $(181/348)$
≥ 6	57.39% $(163/284)$	51.91% $(163/314)$
≥ 7	58.49% $(155/265)$	53.82% $(155/288)$

Although we evaluated our method on all open-class parts of speech with the exclusion of adverbs, we remark that nouns are by far the most frequent case, like in the SemCor corpus (in our random selection of words, they occurred more than half of the times), or the most relevant instances (e.g. when typed as queries to be matched against pages previously tagged with automatic semantic annotations).

Finally, the overall precision beats the baseline by many points in all the experiments, while the difference in recall is not statistically significant. This is a major feature of our approach, enabling precise justifications for sense choices in terms of semantic graphs from which the human validator can benefit in order to take the final decision.

4 Conclusions

In this paper we discussed the use of semantic interconnections to support validators in the difficult task of assessing the quality of both manual and automatic sense assignments. The use of semantic interconnection patterns to support validation allows to smooth possible divergences between the annotators and to corroborate choices consistent with the lexical knowledge base. Furthermore, the method is independent of the adopted lexicon (i.e. WordNet), in that patterns can be derived from any sufficiently rich ontological resource[6].

An interesting point in favour of our approach is that the validator can visually analyse the correctness of a sense choice in terms of its se-

[6] An experiment on the Oxford Dictionary of English is planned in the context of a joint collaboration with Ken Litkowski (CL Research)

mantic interconnections with respect to the other word senses chosen in context. The method has been implemented as a visual tool available online, called *Valido*[7]. The tool takes as input a corpus of documents whose sentences are tagged by one or more annotators with word senses from the WordNet inventory. The user can browse the sentences, and adjudicate a choice over the others in case of disagreement among the annotators.

The tool could be used in the future to collect new, consistent collocations that could grow the lexical knowledge base from which the semantic interconnection patterns are extracted, possibly in an iterative process. We are investigating this topic in an ongoing work.

Moreover, the approach allows the validator to discover mistakes in the lexicon: for instance, the semantic graphs analysed in a number of experiments helped us find out that a *Swiss canton#1* is not a chinese city (*canton#1*) but a division of a country (*canton#2*), that a *male horse* should be a kind of horse, and so on. These inconsistencies of WordNet 2.0 were promptly reported to the resource maintainers, and most of them have been corrected in the latest version of the lexicon.

Finally, we would like to point out the fact that, in the future, semantic interconnections could also be used during the annotation phase by taggers looking for suggestions based on the structure of the lexical knowledge base, with the result of improving the coherence and awareness in the decisions to be taken.

Acknolwedgements

This work is partially funded by the Interop NoE (508011), 6^{th} European Union FP.

References

Agirre Eneko and Rigau German. 1996. Word Sense Disambiguation using Conceptual Density. *Proc. of COLING 1996*, Copenhagen, Denmark.

Bentivogli Luisa, Forner Pamela, and Pianta Emanuele. 1999. Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus In *Proc. of COLING 2004*, Geneva, Switzerland.

Berners-Lee Tim. 1999. *Weaving the Web*, San Francisco, Harper.

Chklovski Tim and Mihalcea Rada. 2002. Building a Sense Tagged Corpus with Open Mind Word Expert. *Proc. of ACL 2002 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, PA.

Decadt Bart, Hoste Véronique, Daelemans Walter, and van den Bosch Antal. 2004. GAMBL, Genetic Algorithm Optimization of Memory-Based WSD. *Proc. of ACL/SIGLEX Senseval-3*, Barcelona, Spain.

Edmonds Philip and Kilgariff Adam. 1998. Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4).

Fellbaum Christiane (ed.). 1998. *WordNet: an Electronic Lexical Database*, Cambridge, MIT press.

Litkowski Ken. 2004. SENSEVAL-3 Task: Word-Sense Disambiguation of WordNet Glosses. In *Proc. of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain.

Mihalcea Rada and Faruque Ehsanul. 2004. SenseLearner: Minimally Supervised Word Sense Disambiguation for All Words in Open Text. *Proc. of ACL/SIGLEX Senseval-3*, Barcelona, Spain.

Mihalcea Rada and Moldovan Dan. 2001. eXtended WordNet: Progress Report. In *Proc. of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

Miller George, Leacock Claudia, Randee Tengi, and Bunker Ross 1993. A Semantic Concordance. In *Proc. 3^{rd} DARPA Workshop on Human Language Technology*, Plainsboro, New Jersey.

Morris Jane and Hirst Graeme. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1): 21-48.

Navigli Roberto. 2005. Semi-Automatic Extension of Large-Scale Linguistic Knowledge Bases. In *Proc. of 18^{th} FLAIRS Conference*. Clearwater Beach, Florida.

Navigli Roberto and Velardi Paola. 2005. Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7).

Snyder Benjamin and Palmer Martha. 2004. The English all-words task. In *Proc. of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain.

Yuret Deniz. 2004. Some Experiments with a Naive Bayes WSD System. In *Proc. of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain.

[7] Valido is available online at: http://lcl.di.uniroma1.it/valido.