Consistent Validation of Manual and Automatic Sense Annotations with the Aid of Semantic Graphs

Roberto Navigli[*]
Università di Roma "La Sapienza"

The task of annotating texts with senses from a computational lexicon is widely recognized to be complex and often subjective. Although strategies like interannotator agreement and voting can be applied to deal with the divergences between sense taggers, the consistency of sense choices with respect to the reference dictionary is not always guaranteed.

In this article, we introduce Valido, a visual tool for the validation of manual and automatic sense annotations. The tool employs semantic interconnection patterns to smooth possible divergences and support consistent decision making.

1. Introduction

Sense tagging 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 **interannotator 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 analyzing 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 article, we present Valido, a tool for supporting the validation of both manual and automatic sense annotations through the use of semantic graphs, particularly of semantic interconnection patterns (Navigli and Velardi 2005).

[*] Dipartimento di Informatica, Università di Roma "La Sapienza," Via Salaria, 113 - 00198 Roma, Italia. E-mail: navigli@di.uniroma1.it.

2. Semantic Networks and Semantic Interconnection Patterns

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 lexical text cohesion (Morris and Hirst 1991), word sense disambiguation (Agirre and Rigau 1996; Mihalcea and Moldovan 2001), and ontology learning (Navigli and Velardi 2004), etc.

Recently, a knowledge-based algorithm for **word sense disambiguation** called **structural semantic interconnections** (SSI, http://lcl.di.uniroma1.it/ssi) (Navigli and Velardi 2004, 2005), has been shown to provide interesting insights into the choice of word senses by providing structural justifications in terms of semantic graphs. Given a word context and a lexical knowledge base (LKB), obtained by integrating WordNet with annotated corpora and collocation resources (Navigli 2005), SSI selects a semantic graph including those word senses having a higher degree of interconnection, according to a measure of connectivity.

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). 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],[1] supported, among others, by the pattern *intersection*#2 $\xrightarrow{part-of}$ *road*#1 $\xleftarrow{kind-of}$ *thoroughfare*#1 $\xleftarrow{kind-of}$ *street*#2. Semantic interconnection patterns are inspired by several works on semantic relatedness and similarity (Rada et al. 1989; Hirst and St-Onge 1998; Mihalcea and Moldovan 2001).

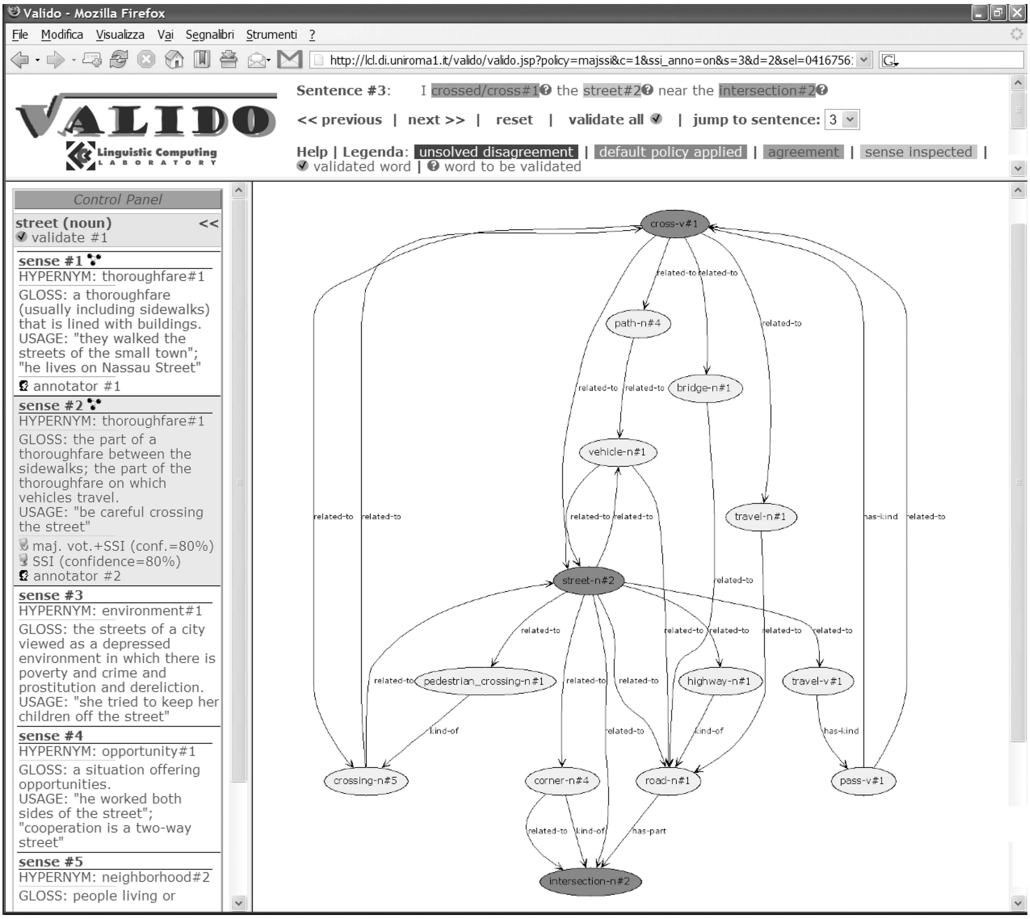
An excerpt of the manually written context-free grammar encoding semantic interconnection patterns for the WordNet lexicon is reported in Table 1. For further details the reader can refer to Velardi 2005.

3. Supporting Validation with Semantic Interconnection Patterns

The validation task can be defined as follows: Let w be a word in a sentence σ, previously annotated by a set of annotators $A = \{a_1, a_2, ..., a_n\}$ each providing a sense for w, and let $S = \{s_1, s_2, ..., s_m\} \subseteq \text{Senses}(w)$ be the set of senses chosen for w by the annotators in A, where Senses(w) is the set of senses of w in the reference inventory (e.g., WordNet). A validator is asked to validate, that is, to adjudicate a sense $s \in \text{Senses}(w)$ for a word w over the others. Notice that s is a word sense for w in the sense inventory, but is not necessarily in S, although it is likely to be. Also note that the annotators in A can be either human or automatic, depending upon the purpose of the exercise.

Based on SSI, we developed a visual tool, Valido (http://lcl.di.uniroma1.it/valido), to support the validator in the difficult task of assessing the quality and suitability of sense annotations. The tool takes as input a corpus of documents whose sentences are

[1] 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 reference inventory. For readability, in the following we omit the noun part of speech.



http://lcl.di.uniroma1.it/valido/valido.jsp?policy=majssi&c=1&ssi_anno=on&s=3&d=2&sel=0416756:

Sentence #3: I crossed/cross#1 the street#2 near the intersection#2

<< previous | next >> | reset | validate all ✓ | jump to sentence: 3

Help | Legenda: unsolved disagreement | default policy applied | agreement | sense inspected |
✓ validated word | ? word to be validated

Control Panel

street (noun) <<
✓ validate #1

sense #1
HYPERNYM: thoroughfare#1
GLOSS: a thoroughfare (usually including sidewalks) that is lined with buildings.
USAGE: "they walked the streets of the small town"; "he lives on Nassau Street"
annotator #1

sense #2
HYPERNYM: thoroughfare#1
GLOSS: the part of a thoroughfare between the sidewalks; the part of the thoroughfare on which vehicles travel.
USAGE: "be careful crossing the street"
maj. vot.+SSI (conf.=80%)
SSI (confidence=80%)
annotator #2

sense #3
HYPERNYM: environment#1
GLOSS: the streets of a city viewed as a depressed environment in which there is poverty and crime and prostitution and dereliction.
USAGE: "she tried to keep her children off the street"

sense #4
HYPERNYM: opportunity#1
GLOSS: a situation offering opportunities.
USAGE: "he worked both sides of the street"; "cooperation is a two-way street"

sense #5
HYPERNYM: neighborhood#2
GLOSS: people living or

Figure 1
Structural interconnection patterns for the sentence *We crossed the street near the intersection* when sense #2 of *street* is chosen, as suggested by the validation policy (γ).

tagged by one or more annotators with word senses from the WordNet inventory. The user can then browse the sentences and adjudicate a choice over the others in case of disagreement among the annotators. To the end of facilitating the user in the validation task, the tool highlights each word in a sentence with different colors, namely, *green* for words having a full agreement, *red* for words where no agreement can be found, and *orange* for those words to which a validation policy can be applied.

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

$S \rightarrow S'S_1 \mid S'S_2 \mid S'S_3$	(start rule)
$S' \rightarrow e_{\mathrm{NOMINALIZATION}} \mid e_{\mathrm{PERTAINYMY}} \mid \epsilon$	(part-of-speech jump)
$S_1 \rightarrow e_{\mathrm{KIND-OF}} S_1 \mid e_{\mathrm{PART-OF}} S_1 \mid e_{\mathrm{KIND-OF}} \mid e_{\mathrm{PART-OF}}$	(hyperonymy/meronymy)
$S_2 \rightarrow e_{\mathrm{KIND-OF}} S_2 \mid e_{\mathrm{RELATEDNESS}} S_2 \mid e_{\mathrm{KIND-OF}} \mid e_{\mathrm{RELATEDNESS}}$	(hypernymy/relatedness)
$S_3 \rightarrow e_{\mathrm{SIMILARITY}} S_3 \mid e_{\mathrm{ANTONYMY}} S_3 \mid e_{\mathrm{SIMILARITY}} \mid e_{\mathrm{ANTONYMY}}$	(adjectives)

A validation policy is a strategy for suggesting a default sense choice to the validator in case of disagreement. Initially, the validator can choose one of four validation policies to be applied to those words with disagreement on which sense to assign:

(α) **majority voting:** If there exists a sense $s \in S$ such that

$$\frac{|\{a \in A \mid a \text{ annotated } w \text{ with } s\}|}{|A|} \geq \frac{1}{2},$$

s is proposed as the preferred sense for w.

(β) **majority voting + SSI:** The same as the previous policy, with the addition that if there exists no sense chosen by a majority of annotators, SSI is applied to w, and the sense chosen by the algorithm, if any, is proposed to the validator.

(γ) **SSI:** The SSI algorithm is applied to w, and the chosen sense, if any, is proposed to the validator.

(δ) **no validation:** w is left untagged.

Notice that for policies (β) and (γ) Valido applies the SSI algorithm to w in the context of its sentence σ by taking into account for disambiguation only the senses in S (i.e., the set of senses chosen by the annotators). In general, given a set of words with disagreement $W \subseteq \sigma$, SSI is applied to W using as a fixed context the agreed senses chosen for the words in $\sigma \setminus W$.

Also note that the suggestion of a sense choice, marked in orange based on the validation policy, is just a proposal and can be freely modified by the validator, as explained hereafter.

Before starting the interface, the validator can also choose whether to add a virtual annotator a_{SSI} to the set of annotators A. This virtual annotator tags each word $w \in \sigma$ with the sense chosen by the application of the SSI algorithm to σ. As a result, the selected validation policy will be applied to the new set of annotators $A' = A \cup \{a_{\text{SSI}}\}$. This is useful especially when $|A| = 1$ (e.g., in the automatic application of a single word sense disambiguation system), that is, when validation policies are of no use.

Figure 1 illustrates the interface of the tool: In the top pane the sentence at hand is shown, marked with colors as explained above. The main pane shows the semantic interconnections between senses for which either there is a full agreement or the chosen validation policy can be applied. When the user clicks on a word w, the left pane reports the sense inventory for w, including information about the hypernym, definition, and usage for each sense of w. The validator can then click on a sense and see how the semantic graph shown in the main pane changes after the selection, possibly resulting in a different number and strength of semantic interconnection patterns supporting that sense choice.

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

3.1 Validating Manual Annotations

In the following, we illustrate the tool by presenting two examples of validation of a manual annotation (the validation policy γ was selected).

Figure 1 shows the senses chosen by the validators for the following sentence:

(a) We crossed the street near the intersection.

Sense #2 of *intersection* and sense #1 of *cross* are marked in green in the top pane, meaning that the annotators fully agreed on those choices. On the other hand, sense #2 of *street* is marked in orange, due to a disagreement between the annotators, one preferring sense #1. Such an inconsistency is reported on the left pane, showing the dictionary definitions of the two senses. The validator can then visualize in the same or in a new window the semantic graphs concerning conflicting sense choices, comparing the interconnection patterns available for sense #1 and #2 of *street*.

After evaluating the respective semantic interconnections, the validator can either confirm the human annotator's choice, accept the SSI interpretation, or assess the semantic interconnection patterns resulting from different sense choices (reported in the left pane of Figure 1).

It is worth mentioning that all the occurrences of the phrase *cross the street* in the SemCor corpus are tagged with the first sense of *street* [defined 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. The tool reflects this fact, showing that both senses are connected with other word senses in context, the first sense having a smaller degree of overall connectivity.[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, 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*). The tool applies the validation policy and suggests sense #6 to the validator.

From these two real-world cases, it is evident that Valido can point at inconsistent, although acceptable, choices made by human annotators due, among others, to the fine granularity of the sense inventory and to regular polysemy. In Section 4 we present an experiment showing that this claim still holds on a larger scale.

[2] In the case of a large, connected graph, a pruned version is shown, and a link is available for viewing a more complete, extended version of the graph.

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.

3.2 Validating Automatic Annotations

While the task of manual annotation is mostly restricted to lexicographers, automatic annotations of texts (especially Web pages) are gaining a huge 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 the fine granularity of the dictionary inventory.

The recognition of intuitive and convincing interconnection patterns reinforces a consistent choice of senses throughout the discourse, a desirable condition for guaranteeing semantic coherence. For example, semantic interconnections can help deal with partially justifiable, but incorrect, interpretations 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 Valido (*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.).

3.3 Weaknesses of the Approach

It can happen that semantic interconnection patterns proposed by the validation tool 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 proposal is provided to the validator, it is less likely that the final choice will be inconsistent with the lexicon structure. Typical examples are:

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

(e) I spent three days in that *hospital*.

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,

Table 2
Precision and recall of the Valido tool in the appropriateness of its suggestions for 360 words.

Part of speech	Precision	Recall	F1 measure
Nouns	73.83% (79/107)	65.83% (79/120)	69.60%
Adjectives	89.29% (25/28)	20.83% (25/120)	33.78%
Verbs	82.14% (69/84)	57.50% (69/120)	67.65%
Total	79.00% (173/219)	48.05% (173/360)	59.76%

apart from the distinction between the act of doing the action and the amount of money paid, there are not many structural suggestions that allow us to distinguish between the two senses. Semantic interconnection patterns cannot help the validator here, but any choice will not violate the structural consistency of the lexicon. As for sentence (e), WordNet encodes two senses for *hospital*: the building where patients receive treatment (sense #1) and the medical institution (sense #2). This case is diametrically opposite in that here WordNet encodes much information about both senses, but such "noisy" knowledge does not help discriminate. As a result, a number of semantic interconnection patterns are presented to the validator, indicating the relevance of both senses for tagging, but no evidence in favor of the choice of sense #1 (which is most appropriate in the sentence).

4. Evaluation

We performed an evaluation of the tool on SemCor (Miller et al. 1993), a selection of documents from the Brown Corpus where each content word is annotated with concepts (specifically, *synsets*) from the WordNet inventory.

The objective of our evaluation is to show that Valido constitutes good support for a validator in detecting bad or inconsistent annotations. A total of 360 sentences of average length (9 or 10 content words) were uniformly selected from the set of documents in the SemCor corpus. The average ambiguity of an arbitrary word in the data set was 5.77, while the average ambiguity of the most ambiguous word in a sentence was 8.70.

For each sentence $\sigma = w_1 w_2 \dots w_n$ annotated in SemCor with the senses $s_{w_1} s_{w_2} \dots s_{w_n}$ ($s_{w_i} \in Senses(w_i), i \in \{1, 2, \dots, n\}$), we identified the most ambiguous word $w_i \in \sigma$, and randomly chose a different sense $\bar{s}_{w_i}$ for that word, that is, $\bar{s}_{w_i} \in Senses(w_i) \setminus \{s_{w_i}\}$. The experiment simulates in vitro a situation in which, for each sentence, the annotators agree on which sense to assign to all the words but one, where one annotator provides an appropriate sense and the other selects a different sense. The random factor guarantees an approximation to the uniform distribution in the test set of all the possible degrees of disagreement between sense annotators (ranging from regular polysemy to homonymy).

We applied Valido with validation policy (γ) to the annotated sentences and evaluated the performance of the tool in suggesting the appropriate choice for the words with disagreement. We assessed *precision* (the number of correct suggestions over the overall number of suggestions from the Valido tool), *recall* (the number of correct suggestions over the total number of words to be validated), and the *F1 measure* $\left(\frac{2pr}{p+r}\right)$.

The results are reported in Table 2 for nouns, adjectives, and verbs (we neglected adverbs, as very few interconnections can be found for them). The experiment shows that evidences of inconsistency are provided by the tool with good precision (and a good F1 measure, especially for nouns and verbs, beating the random baseline of 50%). Notice that this test differs from the typical evaluation of word sense disambiguation tasks, like the Senseval exercises (http://www.senseval.org), in that we are assessing highly polysemous (possibly, very fine grained) words. Comparing the results with a smart baseline, like the most frequent sense heuristic, is not feasible in this experiment, as the frequency of WordNet senses was calculated on the same data set (i.e., SemCor). Notice anyway that beating a baseline is not necessarily our objective if we are not able to provide justifications (like semantic graphs) of which the human validator can take advantage in order to take the final decision.

The low recall resulting for parts of speech other than nouns (mainly, adjectives) is due to a lack of connectivity in the lexical knowledge base, especially when dealing with connections across different parts of speech. This is a problem already discussed in Navigli and Velardi (2005) and partially taken into account in Navigli (2005). Valido can indeed be used as a tool to collect new, consistent collocations that could grow the LKB from which the semantic interconnection patterns are extracted, possibly in an iterative process. We plan to investigate this topic in the near future.

5. Conclusions

In this article we discussed a tool, Valido, for supporting validators in the difficult task of assessing the quality of both manual and automatic sense assignments. The validator can analyze the correctness of a sense choice in terms of its structural semantic interconnections (SSI) with respect to the other word senses chosen in context. The use of semantic interconnection patterns to support validation allows one to smooth possible divergences between the annotators and to corroborate choices consistent with the LKB. Furthermore, the method is independent of the adopted lexicon (i.e., WordNet), in that patterns can be derived from any sufficiently rich ontological resource. Moreover, the approach allows the validator to discover mistakes in the lexicon: For instance, the semantic graphs analyzed 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, that *carelessness* is not a kind of attentiveness, but rather the contrary, 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 that, in the future, the tool could also be used during the annotation phase by taggers looking for suggestions based on the structure of the LKB, with the result of improving the coherence and awareness of the decisions to be taken.

References

Agirre, Eneko and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of COLING 1996*, pages 16–22, Copenhagen, Denmark.

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

Chklovski, Tim and Rada Mihalcea. 2002. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of ACL 2002 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 116–122, Philadelphia, PA.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Hirst, Graeme and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In *C. Fellbaum, editor, WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, pages 305–332.

Litkowski, Kenneth C. 2004. SENSEVAL-3 task: Word-sense disambiguation of WordNet glosses. In *Proceedings of ACL 2004 SENSEVAL-3 Workshop*, pages 13–16, Barcelona, Spain.

Mihalcea, Rada and Dan Moldovan. 2001. eXtended WordNet: Progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100, Pittsburgh, PA.

Miller, George, Claudia Leacock, Tengi Randee, and Ross Bunker. 1993. A semantic concordance. In *Proceedings 3rd DARPA Workshop on Human Language Technology*, Plainsboro, NJ.

Morris, Jane and Graeme Hirst. 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 *Proceedings of 18th FLAIRS International Conference*, pages 548–553, Clearwater Beach, FL, May 16–18, 2005.

Navigli, Roberto and Paola Velardi. 2004. Learning domain ontologies from document warehouses and dedicated websites, *Computational Linguistics*, 30(2):151–179.

Navigli, Roberto and Paola Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7): 1075–1086.

Rada, Roy, Hafedh Mili, Ellen Bickell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.