Evaluating Variants of the Lesk Approach for Disambiguating Words

Florentina Vasilescu, Philippe Langlais, Guy Lapalme
RALI/IRO, Université de Montréal
CP 6128, succursale Centre-ville
{vasilesf, felipe, lapalme}@iro.umontreal.ca

Abstract

This paper presents a detailed analysis of the factors determining the performance of Lesk-based word sense disambiguation methods. We conducted a series of experiments on the original Lesk algorithm, adapted to WORDNET, and on some variants. These methods were evaluated on the test corpus from SENSEVAL2, English All Words, and on excerpts from SEMCOR. We designed a fine grain analysis of the answers provided by each variant in order to better understand the algorithms than by the mere precision and recall figures.

1. Introduction

Many algorithms have been proposed for determining automatically the sense of a word used in a given context. Several word sense disambiguation (WSD) evaluation competitions have been held to score the performance of such algorithms: Senseval 1 and Senseval 2 (Kilgariff & Rosenzweig, 2000; Edmonds, 2002) concentrated on the English language while its European counterpart RomansEval (Calzolari & Corazzari, 2000; Segond, 2000) ran a similar evaluation for French and Italian. Senseval 3 will take place in March 2004, and 90 teams have already declared their interest in participating.

In Senseval 1 and 2, variants of the so-called Lesk approach (Lesk, 1986) were considered either as baseline approaches, or as full fledge systems. In Senseval 1, most of the systems disambiguating English words, were outperformed by a Lesk variant serving as baseline (Kilgariff & Rosenzweig, 2000). On the other hand, during Senseval 2, Lesk baselines were outperformed by most of the systems in the lexical sample track (website).

In the present study, we explore variants of the Lesk algorithm with two goals in mind. First we wanted to evaluate the relevance of the Lesk approach in different settings: on the English All Words Senseval 2 (2473 instances to be disambiguated), as well as on a similar task for Semcor 1.6 excerpts (20964 instances). Due to lack of space, we only report results for Senseval 2, but similar results were obtained for Semcor. Second, we devised a new way of analyzing the results based on a classification of the *risks taken* by each variant in order to better understand the performance of the algorithms at a finer granularity level than one offered by precision and recall.

2. Studied Lesk Variants

The strategy behind all our implementations is to count the number of overlaps between information about *t*, the target word being disambiguated, and its context, i.e. the words surrounding it (we mean by *word*, an open-class lemma). The information about *t* is the bag of words (BOW) representing definitions and/or relations describing each sense associated by WORDNET 1.7.1 to *t*; the information about the context is the BOW containing either the context words or the descriptions of the context words (see section 2.1). The selected sense is the one having the greatest number of overlaps. When there is no overlap between the two bags of words, we select the most frequent sense for *t* as given by WORDNET.

2.1. Context Setting

We distinguish our variants by the way the BOW is associated with the context. Following Lesk (1986), a first class of variants (called OL for Original Lesk) takes into account the descriptions of all senses of the words within the context. In a second class (SL for Simple Lesk) (Kilgariff & Rosenzweig, 2000), the BOW associated with the context is only the context itself, ignoring the sense of its words.

We studied the effect of varying the number of words in the context for the disambiguation. We tested symmetric contexts, centered in the target word, i.e. for $(\pm 2, \pm 3, \pm 8, \pm 10, \pm 25)$ contexts. Audibert (2003) observes that a symmetric context is not optimal in the case of verbs, the most useful information for disambiguation being concentrated on the right side, i.e. on the object of the verb. In section 4, we discuss the influence of the context length upon the performance of our implementations.

On the other hand, Crestan & al. (2003) suggest that automatic context selection would produce better results for certain syntactic categories of words. From this point of view, our variants can be grouped in two categories: one for which all open–class words from the context are taken into account, the other where only the words belonging to the *lexical chain* of the target word are considered. A more detailed description of the way lexical chains are computed is provided in section 2.4.

2.2. Sense Description

We also tested different ways of building the description of a sense. The variant, named DEF (for definition) gathers all the open-class lemmas associated with the definition of the sense provided by the field gloss of WORDNET. When examples are provided, we also consider them as part of the description. The variant named REL (for relation) takes into account the synonyms (synset) of the current sense and all the synsets in a hypernymy relation with the current sense, up to the top of the WORDNET hierarchy. We also tried a combination of these two approaches, named DEFREL, for computing the BOW of a sense.

2.3. Weighting Scores

In the simplest variant, the score assigned to a candidate sense is the number of overlaps between the BOW of that sense and the BOW of the context. Another type of variant, called WHG (for weighted) also takes into account the length of the description for a given sense.

According to Lesk (1986), long descriptions can produce more overlaps than short ones, and thus dominate the decision making process. We multiplied the number of overlaps for a given candidate sense by the inverse of the logarithm of the description length for this sense. Other weighting metrics were also tried, taking into account the distance between a word in the context and the target word, or the frequency of the context word in the language, but that did not bring any significant difference.

2.4. Lexical Chain Selection

Hirst & St-Onge (1998) suggest that words in a text are linked by cohesive relations, forming a *lexical chain.* They use this concept for the detection and correction of *malapropism* (confusion of words having similar pronunciation or spelling, but different meanings). We adapted this idea to word sense disambiguation, selecting within the context only words from its lexical chain. Our implementation (named CL) uses the synonymy and hypernymy relations in WORDNET and a similarity measure, Jaccard formula (Manning & Schütze, 1999), to determine if a given word is a member of the lexical chain.

Given *t* the target word and *w* a word from its context, *Set(t)* is the set of synonyms and hypernyms of all senses of *t* according to the WORDNET hierarchy, and *Set(w)* is the corresponding set of synonyms and hypernyms of all senses of *w*. *w* belongs to the lexical chain of *t* if the Jaccard score computed for *Set(t)* and *Set(w)* is greater than an experimental threshold. An example of this procedure is given in Figure 1, for the context word *legislature,* and the lexical chain of the target word *committee*:

Set(committee) = {committee, commission, citizens, administrative-unit, administrative-body, social-group, group, grouping}

Set(legislature) = {legislature, legislative-assembly, general-assembly, law-makers, assembly, gathering, assemblage, social-group, group, grouping}

> **Committee** approval of Gov._Price_Daniel's "abandoned property" act seemed certain Thursday despite the adamant protests of Texas bankers. Daniel personally led the fight for the measure, which he had watered_down considerably since its rejection by two previous **Legislatures**, in a **public** hearing before the **House_Committee_on_Revenue_and_Taxation**. Under **committee** rules, it went automatically to a **subcommittee** for one week.

Figure 1: Lexical chain of the word *committee* (words in the lexical chain are in boldface)

2.5. Naïve Bayes Approach

We have also implemented a version based on the Naïve Bayes method, choosing from the senses *s* of the word *t* the one which maximizes the quantity *p(s|Context(t)),* making the assumption that there is no dependence between the words in the context of *t*. Our score function is in this case :

$$Score(s) = \log p(s) + \sum_{w \in Context(t)} \log(\lambda p(w \mid s) + (1-\lambda)\, p(w))$$

The three distributions *p(s)*, *p(w|s)* and *p(w)* are computed by relative frequency, using the SEMCOR corpus. The smoothing of *p(w|s)* by a unigram model *p(w)* is controlled by a unique parameter λ, set to 0.95 in our experiments.

3. Evaluation Metrics

Precision and recall ratios provide a quick glance at the merit of each variant but they do not help much in understanding the reasons for their good or bad performance. Therefore we have devised a categorization that distinguishes seven types of answer a Lesk variant can make, according to a reference (gold standard) and to the baseline we use when no overlap is observed. This categorization is better explained by the tree in Figure 2:

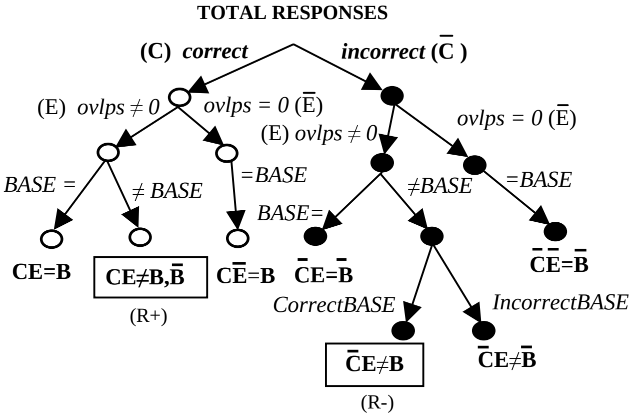
TOTAL RESPONSES
(C) correct
incorrect ($\overline{C}$)
(E) ovlps ≠ 0
ovlps = 0 ($\overline{E}$)
(E) ovlps ≠ 0
ovlps = 0 ($\overline{E}$)
BASE =
≠ BASE
=BASE
≠BASE
=BASE
BASE=
CE=B
CE≠B,$\overline{B}$
C$\overline{E}$=B
$\overline{C}$E=$\overline{B}$
$\overline{C}\overline{E}$=$\overline{B}$
(R+)
CorrectBASE
IncorrectBASE
$\overline{C}$E≠B
$\overline{C}$E≠$\overline{B}$
(R-)

Figure 2: Decision tree of responses produced by a word sense disambiguation system

The tree classifies the responses of an algorithm by comparing a WSD with a reference system BASE, in this case, the one that always chooses the most frequent sense according to WORDNET.

Each node in the tree introduces a binary distinction between answers. The root node discerns whether the answer is correct or not (C = *correct*, $\overline{C}$ = *incorrect*). The second level differentiates the answers when there are overlaps to choose from (we call them *E*, effective answers) from the ones made by default ($\overline{E}$). The third level of nodes separates answers that are identical with the baseline (= *BASE*) and the ones that differ (≠ *BASE*). Finally, the correctness of the baseline is also considered, (B = *correct BASE*, $\overline{B}$ = *incorrect BASE*).

We have defined two new types of measure that we dubbed *risks*. The *positive risk* (*R*+) is determined by the number of correct effective answers that are different from the correct or incorrect baseline ($CE \neq B, \overline{B}$). The case $CE \neq B$ only occurs when the gold standard accepts more than one correct answer for a given target word, and the system and BASE answers are both correct but different. The case $CE \neq \overline{B}$ takes into account the correct effective answers that are different from the incorrect baseline ones. The *negative risk* (*R*-) is determined by the number of incorrect effective answers that are different from the correct baseline ($\overline{C}E \neq B$). We call the difference of these two risks the *gain.* In section 4.2, these measures are given as ratios over the total number of instances processed.

4. Experimental Results

4.1. Comparison of Different Classifiers

The Table 1 presents the precision (P) and recall (R) produced by the variants depicted in section 2, in their DEF version.

%	P ±2	R	P ±3	R	P ±8	R	P ±10	R	P ±25	R
OL	42.64	42.26	42.96	42.58	43.21	42.82	43.29	42.90	42.39	42.01
+***WHG***	39.29	38.94	39.41	39.06	41.21	40.84	40.76	40.40	41.49	41.12
+***CL***	58.38	57.86	58.22	57.70	56.18	55.68	55.65	55.16	53.90	53.42
SL	58.18	57.66	57.20	56.69	54.67	54.19	53.28	52.81	50.47	50.02
+***WHG***	56.67	56.17	55.49	54.99	51.08	50.63	49.25	48.81	44.39	44.00
+***CL***	59.08	58.55	59.12	58.59	58.43	57.91	58.26	57.74	57.41	56.89
BAYES	57.60	57.30	58.00	57.70	56.80	56.60	57.60	57.30	58.50	58.30

Table 1: Precision and recall for different WSD context sizes and DEF description. Underlined figures are the ones that are better than the baseline **P**=57.99, **R**=57.62 (the most frequent sense).

Original Lesk version (**OL**) has a lower performance than the other ones and even than the baseline system. This observation is consistent with Litkowski (2002) hypothesis that only about one third of the instances can rely on the Lesk-style information (definitions and examples) in a disambiguation process. The simplified version (**SL**), in which we only count the overlaps between the description of a candidate sense and the words in the context (and not their description), produces better results in our experiments. Though, improvement of the OL method performances has been observed in the case of POS and sense filtering (see section 4.3). Weighting scores by the inverse of the log of context size does not produce any significant gains, both for **OL** and **SL** methods.

Context selection, as in the case of the lexical chain variant (**CL**), improves the performance for almost all versions. Consequently, not all words of the context seem useful for decision making. It is however surprising that this type of method produces good results for even very small contexts, of ±2 words around the target. The next section proposes a possible explanation of this behavior.

Except for variants **CL** and **BAYES**, increasing the context size lowers performance, which suggests once more the importance of the context selection in the disambiguation process. However the gain of the best variant described in Table 1 is low compared with BASE and we will now try to better understand this phenomenon.

4.2. Response Analysis

Table 2 shows the *positive risk* ($CE \neq B, \overline{B}$) and the *negative risk* ($\overline{CE} \neq B$) incurred by the different variants of SL algorithm. The values are ratios computed according to the total number of processed instances. Table 2 indicates that generally, except for CL methods, the classifiers assume more negative risk than positive one, which reinforces the intuition on the importance of context selection. This tendency becomes more apparent as the context gets longer.

	±2		±3		±8		±10		±25	
%	R+	R-	R+	R-	R+	R-	R+	R-	R+	R-
SL	3.5	3.3	3.9	4.7	6.0	9.3	6.5	11.2	7.8	15.3
+WHG	3.5	4.8	3.9	6.4	5.9	12.8	6.4	15.2	7.8	21.3
CL	1.1	0.2	1.2	0.2	1.7	1.3	1.7	1.5	1.9	2.5

Table 2: Positive risk (R+) and negative risk (R-) for SL variants (negative gains are underlined)

As shown in Figure 3, the systems produce very few answers different from the baseline. Most of the correct answers match the correct baseline, both in the case of effective decisions $CE=B$ and of default answers (most frequent sense) $\overline{CE} = B$. CL correct answers are most often default answers and its success depends on a *silent strategy*, i.e. a few but often correct effective decisions.

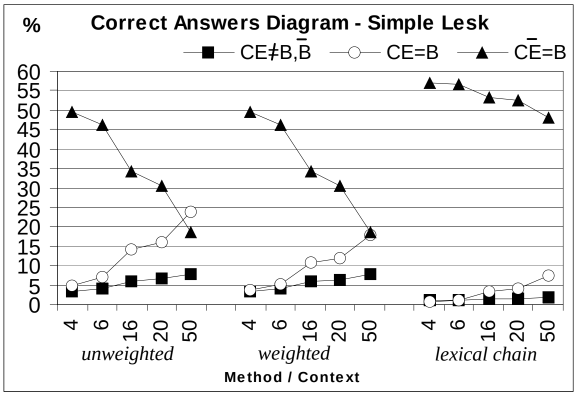

Figure 3:Correct Answers for Simple Lesk Methods

The other variants take more risk and consequently their ratio of correct effective answers different from baseline ($CE \neq B, \overline{B}$) is higher, but increasing the positive risk means that negative risk increases too, as we can also see in Table 1 and Figure 4 below.

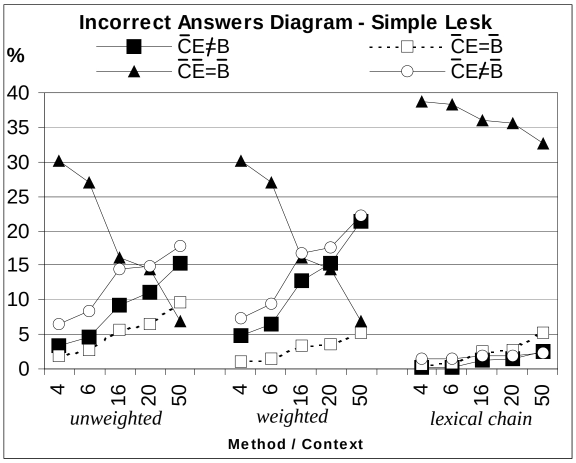

Figure 4:Incorrect Answers for Simple Lesk Methods

An analysis of the incorrect answers (Figure 4) shows that most of the incorrect answers for CL variant coincide with the incorrect baseline ($\overline{CE} = \overline{B}$), which is consistent with the *silent strategy* mentioned before. Thus, there are few effective incorrect decisions taken by this kind of method, and the negative risk ($\overline{CE} \neq B$) is low. On the contrary, the number of incorrect effective decisions taken by the weighted and unweighted methods is higher and so is the negative risk. We can also observe (Figure 3, 4) that the curves for $CE \neq B, \overline{B}$ and $\overline{CE} \neq B$, i.e. for the positive and negative risk, have a similar shape. The small gains in comparison with the baseline could therefore be explained by the difficulty of increasing the positive risk without also increasing the negative one.

4.3. POS and Sense Filtering

We have also studied the impact of the information given by the Part of Speech (POS) on the performances of our system: *a priori* knowing the POS label of the target word (**APOS**), or being able to estimate it using a POS tagger developed in our lab (**RALI**). The *a priori* POS labels have been extracted from the .MRG files provided by the Senseval website. The RALI tagger is a 3rd order Markov model, trained on HANSARD, the Canadian parliamentary debates corpus of a very different kind than the Senseval 2 test corpus. A *POS label* actually behaves as a filter. For example, the word *house* has, according to WORDNET, 12 senses as noun and 2 senses as verb. Knowing (APOS) or guessing (RALI), the POS reduces the number of candidate senses taken into account in the disambiguation process.

We have also obtained better performances by multiplying the score of a candidate sense by a sense-filtering coefficient, calculated according to WORDNET and related to the frequency of a given sense. Table 3 presents the best performances of our system obtained by POS and sense filtering.

SL+CL ±3	**P %**	**R %**	**OL + WHG** ±25	**P %**	**R%**
APOS	61.94	61.34	**APOS + sense filter**	62.52	61.91
RALI	60.48	59.92	**RALI + sense filter**	61.14	60.57
BASE+APOS	61.90	61.30	**BASE+APOS + sense filter**	61.90	61.30
BASE+RALI	60.44	59.88	**BASE+RALI + sense filter**	60.44	59.88

Table 3: Best precision and recall of Lesk approach for sense filtering, known (APOS) and estimated (RALI) POS; BASE: P=57.99, R=57.62.

We can observe that the performances produced by POS and sense filtering are better than the baseline, but still very close to the performance of a baseline also using this kind of filters. Table 3 points out an improved behavior of the OL+WHG ±25 variant, for Senseval2 corpus. However, our experiments on excerpts from Semcor indicate best performances for the OL+CL $\pm 2, \pm 3$ method, in the case of sense and POS filtering. POS, sense filtering and little context lengths have also been reported in Senseval2 exercise, English All Words Task (best supervised system: P=69, R=69; best unsupervised system: P=57.5, R=56.9, as indicated by the Senseval website). According to our study, these factors seem to determine a better behavior of the risk balance.

5. Conclusion

We have implemented and tested several variations of the WSD Lesk algorithm for which we conducted a fine grain analysis of the answers provided by each variant, in order to better understand their performance. We consider this is more informative and productive than just looking at precision and recall figures.

Our experiments have demonstrated that, generally, the performances are lower for larger contexts, the best results being obtained for 4, 6 open-class words contexts around the target word. From this point of view, context selection as in a lexical chain variant seems useful, because in this case the classifier takes less negative risk.

The part of speech information is also an important factor. It diminishes the number of candidate senses of a target word, behaving as a filter. The results obtained by making use of this kind of information and of a sense frequency filter outperform the results of any other classifier that doesn't use these kinds of filtering. The improvement produced by these filters can also be explained in terms of a better risk balance. However the gains obtained if the BASE system also uses this filters, are not very high.

Our study confirms Veronis (2001) conjecture that definitions, examples and relations provided by a dictionary are not sufficient resources in disambiguating words. Other types of information, of a syntactic or pragmatic nature, are needed in order to obtain good performance. Our fine grain analysis in terms of *risks* taken by the WSD algorithms explains why this is so.

References

Audibert, L. (2003). Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurences, TALN, Batz-sur-Mer, 11-14 juin.

Calzolari, N., Corazzari, O. (2000). Senseval/ Romanseval: The Framework for Italian, Computers and the Humanities 34 : 61-78, Kluwer Academic Publishers, Printed in Netherland.

Crestan E., El-Bèze M., de Loupy C. (2003). Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique?, TALN 2003, Batz-sur-Mer, 11-14 juin 2003.

Edmonds, P. (2002). SENSEVAL : The Evaluation of Word Sense Disambiguation Systems, ELRA Newsletter, Vol. 7, No. 3.

Hirst G., St-Onge D. (1998), Lexical Chains as Representations of Context for the etection and Correction of Malapropisms, WordNet an Electronic Lexical Database, MIT Press, (pp. 305-331).

Kilgarriff, A. et Rosenzweig, J. (2000). Framework and Results for English SENSEVAL, Computers and the Humanities, 34, (pp. 15-48).

Lesk M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, ACM SIGDOC '86, The Fifth International Conference on Systems Documentation, Proceedings of ACM Press.

Litkowski, K. C. (2002). Sense Information for Disambiguation: Confluence of Supervised and Unsupervised Methods, Proceedings of the SIGLEX / SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia.

Manning, C.D., Schütze, H. (1999). Foundations of Statistical Natural Language Processing, MIT Press.

Segond, F. (2000). Framework and Results for French, Computers and the Humanities 34 : 49-60, Kluwer Academic Publishers, Printed in Netherland.

Senseval site – http://www.cs.unt.edu/~rada/senseval/

Véronis J. (2001). Sense tagging: does it make sense?, Proceedings of the Corpus Linguistics 2001 Conference, Vol. 13, Special Issue.