Anaphora Resolution with Word Sense Disambiguation

Judita Preiss*
Computer Laboratory
JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
Judita.Preiss@cl.cam.ac.uk

Abstract

We describe a simple word sense disambiguation system equipped with the Kennedy and Boguraev (1996) anaphora resolution algorithm, evaluated on the SENSEVAL-2 English all-words task. The system relies on the structure of the WordNet hierarchy to pick optimal senses for nouns in the text. Since anaphoric references are known to indicate the topic of the text (Boguraev et al., 1998), they may aid disambiguation.

1 Introduction

We investigate the effect of repeating pronominalized nouns in the input to our Word Sense Disambiguation (WSD) algorithm (Preiss, 2001). The WSD algorithm is based on the WordNet 1.7 hierarchy (Miller et al., 1990), and assigns (WordNet) senses to all nouns. The enriched version we evaluate in this paper makes use of our re-implementation of an anaphora resolution algorithm of Kennedy and Boguraev (1996).

If, as claimed by Boguraev et al. (1998), the topic of the discourse is thus repeated, then the main topic words will be more likely to be disambiguated correctly. The subsequent WSD algorithm makes use of this extra topic information, and this will in turn affect the disambiguation of all other nouns in the discourse.

The system is evaluated on the English all-words task in SENSEVAL-2.

2 Algorithms

2.1 Overview of the Algorithm

Our WSD algorithm has three components, as depicted in Figure 1. Taking as input the test data parsed using the Briscoe and Carroll (1993) parser (which uses the grammar described in Carroll and Briscoe (1996)), the first step is to identify and discard the pleonastic pronouns. Our pleonastic component is described in section 2.2.

In the next phase (section 2.3), third person pronouns are resolved to a noun antecedent and replaced in the text by the noun antecedent. The purpose of this is to increase the number of topic words in the text, to aid the disambiguation of other nouns. This approach assumes firstly that pronouns refer mainly to topic words, and secondly that repeating topic words in the text helps overall disambiguation.

The final phase of the algorithm is the WSD component, described in section 2.4. Using simulated annealing, it attempts to find a sense assignment for every noun that minimizes an overall 'distance' function using the WordNet hierarchy. In addition, for the repeated nouns added in the previous phase, the senses are tied together. This means that if the sense of one word in a tie is changed during simulated annealing, the sense of all words in the tie are simultaneously changed.

The advantage of this approach can be shown on the following discourse: *The parrot, like the chicken, is kept by people as a domesticated bird. It can speak.* Suppose firstly that there is no anaphora resolution phase. The words *parrot, chicken, person, bird* are given to the word sense disambiguation algorithm, and the system chooses senses which are related to people (*parrot* in the sense of mimicking people, *chicken* a wimp and so on). This is clearly incorrect. Now suppose we resolve the pronoun *it* to *parrot*, and repeat the word *parrot* in the text. Now the words *parrot, chicken, person, bird, parrot* are passed to the WSD system (where the two

* This work was supported by the EPSRC while the author was at the University of Sheffield.

parrots are sense-tied together), and the system now chooses the correct bird-related senses.

2.2 Pleonastic Pronouns Component

It can be a pleonastic pronoun (pronoun with no antecedent), for example in the sentence: *It is raining.* We label the pronoun *it* as pleonastic if it is a subject of a raising verb (these were extracted from the ANLT lexicon (Boguraev and Briscoe, 1987)) or if it was used in conjunctions with the verb *to be* and one of a particular set of adjectives (for example *It is possible to go to town.*).

The component was evaluated on a manually anaphorically resolved portion of the BNC (the initial 2000 sentences of w01). It has a precision (proportion of pronouns deemed pleonastic which really are pleonastic) of 94% and recall (proportion of pleonastic pronouns recognized as pleonastic) of 61%.

2.3 Anaphora Resolution Component

The pronominal anaphora resolution is carried out by our re-implementation of the Kennedy and Boguraev (Kennedy and Boguraev, 1996) anaphora resolution algorithm. This algorithm is based on that of Lappin and Leass (Lappin and Leass, 1994), but does not require a full parse. It treats the cases of third person pronouns and lexical anaphors.[1] Its cited accuracy is 75% on general corpora (Kennedy and Boguraev, 1996), but note that their published algorithm uses the LINGSOFT morphosyntactic tagger.

The algorithm creates coreference classes which join together words which are believed by the algorithm to be referring to the same object. These classes are assigned a salience value based on the presence of the features in Table 1. The salience value of a class is the sum of the feature weights of its members, scaled down by the number of sentences ago that the feature last occurred. The correct antecedent is chosen to be the closest word from the coreference class with the highest salience.

2.4 WSD Component

We define a notion of distance between any two WordNet noun senses which is based on the WordNet hierarchy.[2] As pointed out by Resnik (1999), it is naive to assume that the distance between any two nodes in the hierarchy is equal. We therefore assign a weight w to every noun sense x:

$$\text{weight}(x) = \frac{\text{number of children below } x \text{ in hierarchy}}{\text{total nodes in hierarchy}}$$

This is used to define the distance between two distinct noun senses x and y:

$$\text{dist}(x,y) = \min_{z \in h(x) \cap h(y)} \text{weight}(z) - \tfrac{1}{2}\text{weight}(x) - \tfrac{1}{2}\text{weight}(y)$$

where $h(s)$ denotes the hypernym chain of noun sense s.[3] If the hypernym chains of x and y do not intersect, the distance is set to the maximum value of 1. In Preiss (2001), we investigated scaling the distance function such that for noun senses x and y at positions in the corpus n and m respectively:

$$\text{dist}^*(x,y) = \frac{\text{dist}(x,y)}{|n-m|^\alpha}$$

Note that we do not explicitly use a window of surrounding nouns, but the $|n-m|$ denominator means that contributions from far away nouns are usually negligible. We showed that it was not possible to guess the optimal value of α

Condition	Weight
Current sentence	100
Current context	50
Subject	80
Existential construct	70
Possessive	65
Direct object	50
Indirect object	40
Oblique	30
Non embedded	80
Non adjunct	50

Table 1: Salience values

[1] Lexical anaphors are reflexives and reciprocals.

[2] In the SENSEVAL-2 task we identify nouns by using an enhanced version of the GATE tagger and lemmatizer (Cunningham et al., 1995).

[3] The hypernym chain of s consists of the word s, the parent word of s, the grandparent of s, etc, all the way to a root word.

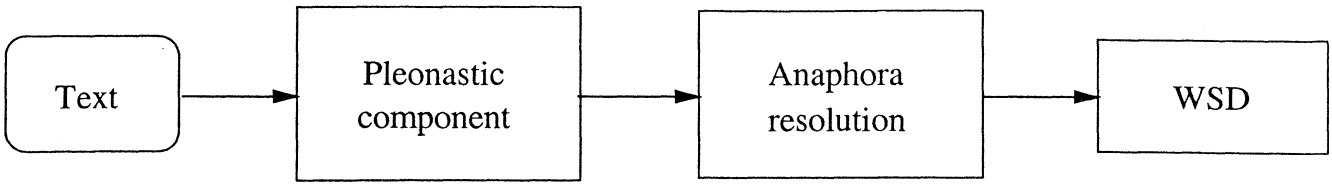

Figure 1: Integration of components

in advance for any set of texts covered in SEMCOR. However, averaged over all words there is a slight peak around $\alpha = 1$, so this is the value we take.

The distance between two adjacent nodes in the hierarchy may now not be equal. To illustrate this, consider the following example adapted from a paper of Resnik (1999). In WordNet 1.7 (prerelease), VALVE is the parent node of SAFETY VALVE, and MACHINE is the parent of INFORMATION PROCESSING SYSTEM. However, the intuitive distance between the first pair of nodes seems to be less than the distance between the second pair. Using our distance function outlined above, the distance between SAFETY VALVE and VALVE is 0.000121, while the distance between INFORMATION PROCESSING SYSTEM and MACHINE is 0.00229. This is depicted in Figure 2.

We want to assign precisely one sense to each noun in the text; we call this a path. We find the 'optimal' path by simulated annealing (Bertsimas and Tsitsiklis, 1992). Simulated annealing is a probabilistic method for finding the global optimum of a function which may have a number of local optima. We define the function to be minimized, the energy function, to be the sum of all the pairwise scaled distances.

Our version of simulated annealing starts with a randomly chosen path which it attempts to improve. It performs a number of iterations in which it randomly chooses a word and then chooses a new sense for this word.[4] If this change is an improvement in terms of the energy function, it is kept. Otherwise, it may or may not be accepted depending on the current value of the temperature. Over time the temperature decreases, making it less likely to keep changes that increase the energy. The algorithm terminates when no changes were made in the last 1000 iterations.

When simulated annealing terminates, it outputs what it deems the optimal sense assignment for all the nouns in the text. For a more detailed description of the WSD algorithm, please refer to Preiss (2001).

This algorithm was implemented in C and executed on a Pentium III 500MHz. Each text took 1 hour to initialize, and 2 hours to perform 20 runs of simulated annealing. A majority vote then decided the sense assignment.

Article	Words	Senses	Ties
1	363	1698	38
2	575	2098	46
3	340	1495	60

Table 2: Test data for the English all words task

3 Results

The WSD component enhanced with the anaphora resolution algorithm was submitted for the English all-words task in SENSEVAL-2. The test data for this task consisted of three articles, and information gathered from each article is displayed in Table 2. The words column shows the number of words marked as nouns by the part of speech tagger in the parser. The senses column contains the total number of senses for all of these words. The ties column shows the number of ties in the text, where each tie contains a noun and some pronouns that refer to it. The system achieved 44% precision and 20% recall fine-grained, and 45.2% precision and 20.5% recall coarse-grained.[5]

[4] We slightly skewed the probability distribution of the senses towards the more frequent sense. The probability of the nth sense is proportional to $\frac{1}{n^2}$.

[5] The system assigns senses to all nouns but to no other part of speech. It also has no mechanism for marking a word undecidable.

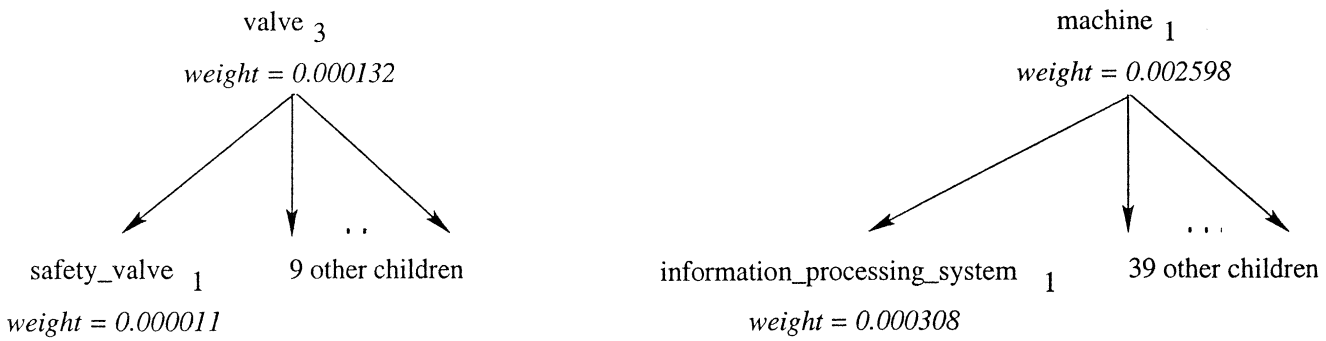

Figure 2: Distance between adjacent nodes

4 Future Work

We would like to investigate the performance of the WSD system with and without anaphora resolution, with a view to also extending links in text to other entities.

Although the precision of the pleonastic component is currently quite high, we intend to boost recall possibly by including some of the rules devised by Lappin and Leass (1994).

Acknowledgements

I would like to thank John Carroll for parsing the SENSEVAL-2 corpus for me.

References

D. Bertsimas and J. Tsitsiklis. 1992. Simulated annealing. In *Probability and Algorithms*, pages 17–29. National Academy Press, Washington, D. C. .

B.K. Boguraev and E.J. Briscoe. 1987. Large lexicons for natural language processing: utilising the grammar coding system of the *longman dictionary of contemporary english. Computational Linguistics*, 13(4):219–240.

B. Boguraev, C. Kennedy, R. Bellamy, S. Brawer, Y. Y. Wong, and J. Swartz. 1998. Dynamic presentation of document content for rapid on-line skimming. In *Proceedings of AAAI Spring Symposium on Intelligent Text Summarisation*, pages 118–128.

E. Briscoe and J. Carroll. 1993. Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics*, 19(1):25–60.

J. Carroll and T. Briscoe. 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. In *Proceedings of the ACL SIGDAT Conference on Empirical Methodsin Natural Language Processing*, pages 92–100.

H. Cunningham, R. Gaizauskas, and Y. Wilks. 1995. A general architecture for text engineering (GATE) — a new approach to language R&D. Technical Report CS–95–21, University of Sheffield.

C. Kennedy and B. Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113–118.

S. Lappin and H. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

G. Miller, R. Beckwith, C. Felbaum, D. Gross, and K. Miller. 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

J. Preiss. 2001. Local versus global context for WSD of nouns. In *Proceedings of CLUK 4*, pages 1–8.

P. Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.