1 Corpus-Based Stemming [1]

1.1 Objective:

Common stemmers (e.g. Porter Stemmer) produces results that are too aggressive. E.g. race: {racial, racially, racism, racist, racists}, {races, racing, racer, racers, racetrack}.

This research aims at reducing variant word forms to common roots, so as to improve the precision of an information retrieval system.

1.2 Methodology:

- Find initial equivalent class by an aggressive stemmer.
- Score any pair of the words in the original equivalent class with a similarity value derived from a large corpus.
- Use "Connected Component Algorithm" and "Optimal Partition Algorithm" to find new better equivalent classes.

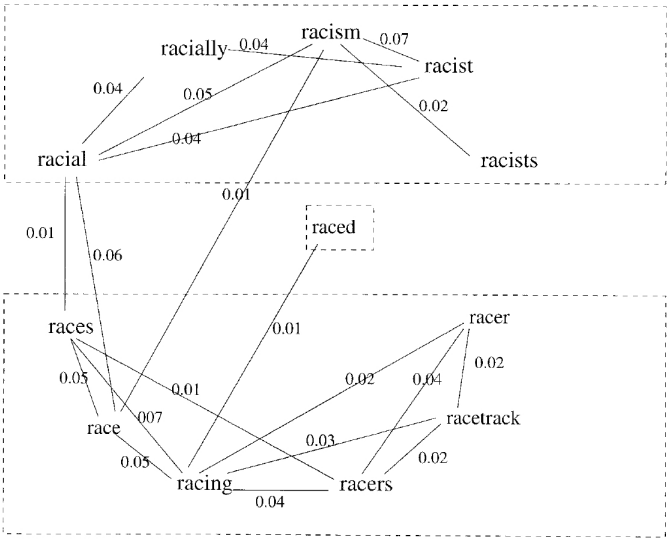

Figure 1: Optimal partition of a connected component equivalence class

1.3 Experiment & Results

- Corpora used for training and testing included WEST legal document collection, and WSJ(87-91) and WSI91(91) from TREC.
- Results show that corpus-based analysis of word variants can be used to enhance the performance of stemming algorithm.

2 Corpus-Based Machine Translation [2]

2.1 Objective:

Achieve machine translation by using statistics of bi-lingual text corpus.

2.2 Methodology:

Define S to be certain text in source language and T to be the text in target language that is observed. Machine translation from T to S can be viewed as the problem of finding certain text S, such that among all the text in the source language S has the highest probability of being translated into T.

Find S and T to maximize $Pr(S \mid T) = \frac{Pr(S)Pr(T|S)}{Pr(T)}$

- $Pr(S)$ can be estimated by using a tri-gram model in source language.
- And $Pr(T \mid S)$ can be estimated by the expression of $Pr(n \mid e) \times Pr(f \mid e) \times Pr(i \mid j, l)$.

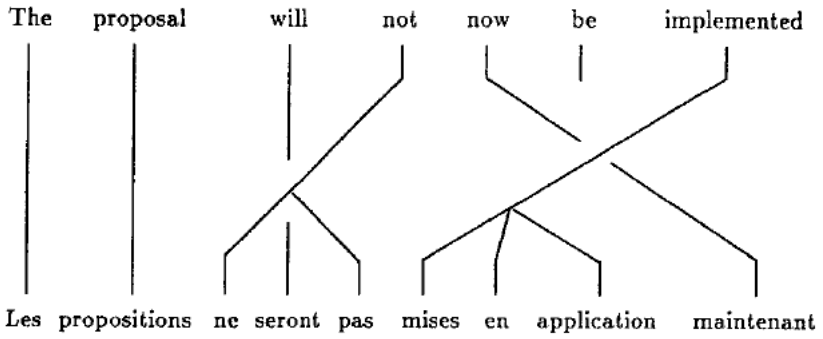

Figure 2: Example Translation

Parameters of probabilities need to be derived from a large bi-lingual corpus.

2.3 Experiment & Results

- Bi-lingual corpus used was the proceedings of the Canadian parliament (100 million words of English text and the corresponding French translation).
- 73 French sentences tested, 5% exactly correct translation, 48% of the translations are acceptable.

3 Corpus-Based Parsing [3]

3.1 Objective:

Build a self-learning parser that may extend itself without relying on extra input from the outside world.

3.2 Methodology:

- Collecting partial results and generating hypotheses based on universal constraints and the parser's current knowledge.

(1) AP(3-11) :- NP(3-5), S(6-11).
(2) NP(3-11) :- NP(3-5), S(6-11).
(3) VP(2-11) :- is(2-2), NP(3-5), S(6-11).
(4) NP(1-6) :- S(1-5), NP(6-6).
(5) S(1-11) :- S(1-5), S(6-11).
(6) Smaj(1-11) :- S(1-5), S(6-11).

Figure 3: An example of the hypotheses generated for the sentence "Lead is a soft metal that serves rnany purposes in home"

- For each set of hypotheses generated for parsing a single sentence, the one that was generated the most of times wins.

3.3 Experiment & Results

- WSJ Corpus was used for verifying the validity of this method.

4 Corpus-Based Word Sense Disambiguation [4]

4.1 Objective:

Have a system learn to disambiguate the appearance of a word W using the appearances of W in an untagged corpus as examples.

4.2 Methodology:

- Using the definition of each entry of a Machine Readable Dictionary (word sense), compute the closely related sentence context.
- Compute the similarities of the context of an appearance of the word W (needs to be disambiguated) with each trained context of a word sense.
- The word sense of the context with highest similarity wins.

4.3 Experiment & Results

- Disambiguation of four noun words (drug, sentence, suit, player) was tested, totally 500 occurrences. Average success rate on the 500 appearances was 92%.
- Testing sentences were chosen from the Treebank-2 corpus.
- Used a combination of the online versions of the Webster's and the Oxford dictionaries, and the WordNet system. WordNet was found to be the single best source of seed words.

5 Corpus-Based Tagging [5]

Brill Tagger, discussed in class with details.

References

[1] Xu, J. and Croft, W. B. (1998). Corpus-based stemming using cooccurrence of word variants, ACM Transactions on Information Systems, 16(1), 61-81.

[2] Brown, F. P., et. al. (1990). A Statistical Approach to Machine Translation, Computational Linguistics, 16(2), 79-85.

[3] Liu, R. and Soo, V. (1994). A corpus-based learning technique for building a self-extensible parser, in Proceedings of the 15th conference on Computational linguistics, 1, 441-446.

[4] Karov, Y. and Edelman, S. (1998). Similarity-based word sense disambiguation, Computational Linguistics, 24(1), 41-59.

[5] Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging, Computational Linguistics, 21(4), 543-565.