SENSE DISAMBIGUATION in the PANGLYZER

Steve Helmreich

Computing Research Laboratory

ARPA Machine Translation Evaluation Workshop

March 17-18, 1994

PANGLOSS

Two Problems of Sense Disambiguation

1. Determining the Standard

2. Determining the Method

Approaches to Problem One

a. Disambiguation to a task

b. Disambiguation to a dictionary

c. Disambiguation to an ontology

d. Disambiguation to internal standard

PANGLOSS

Approaches to Problem Two

1. (explicit) Knowledge-Based

2. Statistical (implicit Knowledge-Based)

    a. Definition overlap (e.g., Cowie, Guthrie and Guthrie)

    b. Probabilistic (e.g., Brown et al.)

    c. Word Space (e.g., Schütze)

PANGLOSS

WORD SPACE ALGORITHM

1. Collect 4-grams from large quantities of text
2. Reduce the number of 4-grams to a reasonable amount
3. Construct a collocation matrix
4. Perform SVD (Singular Value Decomposition) on the matrix
5. Take the top 97 dimensions
6. For every word, create a context vector for every occurrence of that word
7. Normalize and sum the context vectors to create a word (or confusion) vector
8. Cluster the context vectors to produce senses

PANGLOSS

SINGULAR VALUE DECOMPOSITION (SVD)

Given any matrix, A, decompose it into a product of three matrices, U, Σ,V^T, producing a decomposition $A=U\ \Sigma V^T$ such that Σ is a diagonal matrix with decreasing positive values down the diagonal.

By cutting off the lower end of the matrix, Σ, a lower-dimensional matrix than the original is obtainable, that is the best approximation (at that rank) for the original matrix.

PANGLOSS

CRL's APPROACH

Mark Casper and Jim Hargrave

1. Use syllables instead of 4-grams

2. Use sentences as context window

3. Disambiguate sense clusters to LDOCE-WordNet senses

4. Use LDOCE definition and WordNet synset as context vector

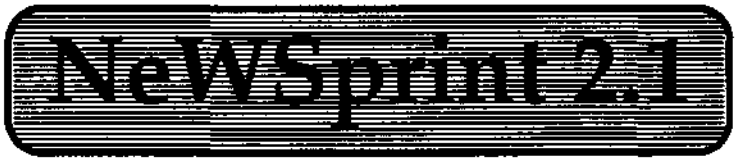
PANGLOSS

shelmrei@ithaka **stdin**

tido

Fri Mar 4 14:24:23 1994

NeWSprint 2.1 Rev B
Openwin library 3
NeWSprint interpreter 3.010



Lessons Learned

- There will be a "VIASA" article in each evaluation
- 70% to 60% in 6 months: we'll be done in 3 years!

Seriously, though:

- Study of inter-subject time variations: who is helped more?
- Incremental test-and-development cycle very important
- Multiple-Engine MT via chart manager looks very promising
- Methodological issue: fully investigating one MT technique versus finding the best-performance combination in the short run

Research Plans

Continued development of current MT engines and chart manager, including:

- Semantic mapper, sentence planner; later, higher-level semantic processing
- Example analysis problem: Ranking techniques for Spanish analysis
- Example generation problem: Article insertion in English generation
- Example of expansion of Pangloss: Japanese/English translation