Word Sense Disambiguation: Pipelining with Supervised LSA

Kiran Vodrahalli, Evelyn Ding, Albert Lee

May 8, 2014

1 Introduction

1.1 Problem Statement

Multiple word senses

ex. bank: "financial institution" or "river bank" or "count on something happening" or "supply/stock held in supply for future use" ...

Wrong sense can drastically alter meaning

Objective: given context, determine which sense of the word is intended

First proposed by Weaver in 1949 as computational task in the early days of MT

One of the oldest open problems in computational linguistics

AI-Complete Problem

Example

"Paintings, drawings and sculpture from every period of art during the last 350 years will be on display ranging from a Tudor portrait to contemporary British art." vs. "Through casual meetings at cafe's, the artists drew together to form a movement in protest against the waste of war, against nationalism and against everything pompous, conventional or boring in the art of the Western world."

the creation of beautiful or significant things vs. the products of human creativity; works of art collectively

2 Previous Work

knowledge based

unsupervised corpus-based

supervised corpus-based

2.1 Knowledge-based

Lesk's algorithm: two words W1 and W2 with multiple word senses for each sense i in W1 for each sense j in W2 calculate overlap of i and j in dictionary definitions maximize overlap i and j to determine word senses Semantic similarity: calculate path between words in wordnet for each word sense with other words in the context minimize total semantic path to select word sense context can vary from one sentence to length of document

2.2 Unsupervised Corpus-based

Token-based discrimination: cluster contexts in which a given target word appears Cluster 1: The line was occupied. The operator came into the line abruptly.

Cluster 2: The line was really long and it took forever to be served. I stood in the queue for about 10 minutes

Use parallel corpus across language to infer sense distinctions across the language Je vais prendre ma propre dcision. vs. Je vais prendre ma propre voiture.

2.3 Supervised Corpus-based

Use annotated corpus to build model ex. to know: 1) be aware of piece of information I want to know who is winning. I know that the president lied. 2) know person She doesn't know the composer. Do you know my sister? Use context of how word usually appears (bigram/trigram models) if (feature) then word sense

2.4 LSA – Past Efforts

3 Data

4 Algorithms

Part of speech tagging and bigram models Use for subset of data that has high predictability Supervised LSA for the rest of the data

Other methods had low accuracy

4.1 Part-of-Speech

4.2 Semantic Knowledgebase

4.3 Supervised LSA

5 Results

6 Conclusions

7 Future Work