Word sense induction using Spark

Stella Lau

28 November 2017

Background

- Word senses are essential in information retrieval, machine translation, and word sense disambiguation
- **Word sense induction**: task of identifying different senses of a target word in a given text
- Vector-based approaches: contexts of a word represented as a vector of features; cluster context vectors
- Graph-based methods
  - Vertices: co-occurring words in contexts of target
  - Edges (v_1, v_2): v_1 and v_2 co-occur in one+ contexts of target
  - Graph-clustering to induce senses

Problem with graph clustering: single senses [1]

Which sense of **network** does system belong to?

> *To install our satellite system please call our technicians and book an appointment. Connection to our television* **network** *is free of charge*
>
> *To connect to the BT* **network**, *proceed with the installation of the connection software and then reboot your system*

Approach: Hierarchical clustering

Graph-based method

- Vertices: collocation[1] with target word
- Edges: weighted based on co-occurrence frequencies of associated collocations

Input

- Base corpus: paragraphs containing target word
- Large reference corpus (e.g. British National Corpus)

Output: sense of the target word

Lots of pre-processing and steps...

[1] A collocation is a juxtaposition of words within the same paragraph

WSI induction algorithm

1. Corpus pre-processing: POS-tag corpora, only keep nouns. Filter noisy nouns
2. Extract collocation, assign weight; smooth to discover new edges
3. Cluster graph using Markov Clustering or similar community detection algorithm
4. Evaluation: SemEval 2007 WSI task

Investigation

Parallel implementation

From input to output, what parts of NLP task can be parallelised?
Is there any advantage of doing so? What is the bottleneck?
Can we generalize?

Distributed graph clustering or community detection

How do we implement distributed graph clustering using GraphX?

Evaluation

Can we do better than the base paper on SemEval 2007 WSI?

Steps

1. Download tasks and corpus data
2. Implement pre-processing algorithm in Spark
3. Experiment with graph clustering in GraphX
4. Evaluate with SemEval WSI task; try to improve
5. Generalize: what would a NLP library in Spark look like?

References I

Ioannis P Klapaftis and Suresh Manandhar. "Word Sense Induction Using Graphs of Collocations." In: *ECAI*. 2008, pp. 298–302.