Domain Adaptation for Paraphrasing

Abstract

Paraphrases are generally not fully applicable to a specific domain of text. Direct paraphrase extraction from a domain-specific typically leads to poor phrase coverage and low quality paraphrases due to the lack of in-domain data. In this paper, we adopted the data selection method introduced by Moore and Lewis to subsample data relevant to the in-domain text from a large non-domain specific corpus based on the difference in cross entropy. Empirical results (biology text?) demonstrate that the paraphrase generated from subsampled corpus using such metric are more relevant to the domain of interest. ¡Data on biology textbook¿ ¡Mturk results¿

1 Introduction

2 Related Work

Frustratingly Easy Domain Adaptation, Hal Daume III, 2007 ? fully supervised setting for learning a function to map input space X to decision/output space Y that optimizes its performance on target (in-domain) data ? the goal of our domain adaptation is to expand target domain data to train some *other* system/model; no explicit output space Y for the paradigm introduced by Hal

3 Paraphrase Extraction

3.1 Paraphrase Acquisition from Bitexts

3.2 Monolingual Distributional Similarity

4 Domain Adaptation

4.1 Data Selection based on Difference in Cross Entropy

4.2 Domain Adaptation for Paraphrasing

Since there are two stages in our paraphrase framework that depends on data statistics, we investigated the effectiveness of paraphrase biasing based on the application of domain adaptation in each of the stages.

Bilingual Pivoting

Monolingual Distributional Similarity

Previous similarity score based on monolingual distributional similarity was computed using the Google ngram corpus (Lin et al., 2010), which comprises of statistics from general domain of resources from the internet. Although more data leads to a better vector space model for the score, reranking of the paraphrases can potentially benefitted from adapting the training corpus to the target domain. testttt

5 Experimental Setup

@@ diagram from description sent to Peter?

5.1 Data and Language Model Parameters

@@ data: in-domain, out-domain @@ srilm model params

5.2 Paraphrase Extraction

@@

6 Experimental Results

6.1 Language Model Training for In-Domain Text

6.2 Paraphrase Examples

7 Conclusion

References

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New tools for web-scale n-grams. In *Proceedings of LREC*.