Creating a Textual Entailment Dataset Automatically from Guardian Articles

Daoud Clarke

August 8, 2011

1 Introduction

Recognising textual entailment is the task of determining, given two sentences, whether the first entails or implies the second. An entailment dataset is a set of triples (t, h, e) where t is the *text* or entailing sentence, h is the *hypothesis* or entailed sentence, and e is a Boolean value indicating whether entailment holds or not. Datasets are normally constructed by manual analysis, which is time consuming and thus limits the size of dataset that can be constructed.

In this document, we will describe a process for constructing a textual entailment dataset automatically, based on an idea introduced by Burger and Ferro (2005) and developed by Hickl et al. (2006).

References

J. Burger and L. Ferro. Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 49–54. Association for Computational Linguistics, 2005.

A. Hickl, J. Williams, J. Bensley, K. Roberts, B. Rink, and Y. Shi. Recognizing textual entailment with lccs groundhog system. In *Proceedings of the Second PASCAL Challenges Workshop*, 2006.