Almost-Unsupervised Cross-Language Opinion Analysis at NTCIR-7

NLCL group: Taras Zagibalov (T.Zagibalov@sussex.ac.uk) John Carroll (J.A.Carroll@sussex.ac.uk) Department of Informatics University of Sussex UK

Summary

We describe the Sussex NLCL System entered in the NTCIR-7 Multilingual Opinion Analysis Task (MOAT). Our main focus is on the problem of **portability** of natural language processing systems **across languages**. Our system was the only one entered for all four of the MOAT languages, **Japanese, English, and Simplified and Traditional Chinese**. The system uses an almost unsupervised approach applied to two of the sub-tasks: **opinionated sentence detection** and **topic relevance detection**.

Our Approach

1. Lexical Item Extraction

Find words or word combinations (Lexical Items) in a language-independent manner. Overcome the problem of word detection in unsegmented texts in Asian languages.

2. Relevance Classification

Find Lexical Items that are relevance indicators by comparing their frequency-based ranks for different topics.

3. Subjectivity Classification

Automatically generate the list of pairs of associated words. Manually filter the lists to find the subjectivity markers for Subjectivity classification of the relevant sentences. Automatically expand the lists to improve performance.

Methods

Lexical Item Extraction

Any sequence of characters that occurs at least three times is a candidate to be a LI

If the frequency of a LI is the same as that of a shorter sub-unit then the latter is deleted.

Relevance Classification

All LI are ranked according to their frequency in a text.

LI frequencies are compared across all the texts.

LI with the biggest rank differences are the relevance indicators.

LI	Topic 1 rank	Topic 2 rank	Difference	Relevance Marker?
the	2	3	1	X
netscape	0	10	10	√
law	24	6	18	√

Subjectivity Classification

For each LI we found immediate neighbours:

第五次缔约方大会的中国代表团

中国： 的 _0, 大会的 _0, 代表团 _1

Headwords	中国	美国	经济	的
Context words	经济	经济	中国	经济
Context words	跟	对	的	快速

RunID1: Use the list of manually filtered words from the automatically generated word lists

RunID2: RunID1 + (χ^2 >average)

RunID3: RunID1 + (χ^2 >3.84)

Results