



LIPN Participation to TAC 2008 : Opinion Summarization Task



Laboratoire d'Informatique de Paris Nord
Université Paris 13 et CNRS UMR 7030
99 avenue Jean-Baptiste Clément
93430 Villetaneuse
FRANCE

Aurélien Bossard
Michel Génèreux
Thierry Poibeau

Michel Génèreux's System (using snippets)

System Overall Architecture

The system is not without resemblance to MEAD. It consists in :

- Cleaning data (1)
- Sentence scoring (2)
- Reranking (3)

Sentence scoring (2)

To compute a score, the system combines different features :

- Similarity with target
- Similarity with query
- Similarity with snippets
- Similarity with first sentence
- Centroid : sum of words' tf.idf
- Position : $\sqrt{1/d(top)}$

A sentence score is the **sum of all these features**.

In addition, the system has been adapted to opinion task , using a new feature (4).

Adaptation to opinion task (4)

A feature has been added to the system to adapt it to « Opinion summarization task ».

A **SVM classifier** has been developed, classifying blogs as positive or negative.

The SVM has been trained on « Pang » corpus.

Each **sentence** is attributed the **same polarity** as the **blog post itself**.

Sentences are grouped together by polarity. The idea is to improve summaries by selecting sentences which opiated polarity is the same as the query

Reranking (3)

In order to **avoid redundancy**, sentences go through a reranking phase : starting from the highest scored sentence, the reranker only inserts a sentence to the summary if it is not too similar to a previously inserted sentence.

Sentence similarity

Two similarity measures have been tested :

- Cosine (bag of words)
- Levenhstein distance : the more operations (word deletion, insertion, replacement) needed, greater the distance

These measures have been tested on Microsoft Research Paraphrasing Corpus. There is no proof of better detecting paraphrases using Levenhstein distance or lexical databases.

Cosine similarity measure has been consequently integrated to our summarization system.

Results

This system obtained the following scores :

- **Pyramid** : 0.393 (best : 0.534, worst : 0.101)
- **Grammaticality** : 6.636 (best : 7.545, worst : 3.545)
- **Non-redundancy** : 6.818 (best :8.045, worst : 4.364)
- **Structure/Coherence** : 3.04 (best : 3.591, worst : 2.)
- **Fluency/Readability** : 4.591 (best :, worst :)
- **Responsiveness** : 4.500 (best : 5.773, worst : 1.682)

With all scores attributed the same weight, the system ranks **first** of all runs.

Cleaning data (1)

This step is common to Michel Génèreux's and Aurélien Bossard's system.

Sentences with ratio
 $card(frequent\ words) / card(total\ words)$
below 0.35 are deemed too noisy and discarded.

Aurélien Bossard's System (not using snippets)

Hypothesis

In multi-document summarization problematic, the most important ideas/facts (in our case, opinions) are those which appear most often in the documents.

We assume sentences contain only one opinion. Grouping sentences by ideas can give us a view of what is important and what has to be extracted.

System Architecture

Starting from our hypothesis, the system must have two phases : one phase for grouping sentences, and one for extracting the ones considered as most important.

- Cleaning (1)
- Sentence similarity computation (2)
- Sentence semantic grouping (Clustering) (3)
- Sentence selection (4)
- Reranking (4)

Clustering (3)

In order to detect sentences conveying the same opinions, we cluster the sentences using their similarity between each other as feature.

The system is using **fast-global kmeans**, an unsupervised clustering method.

Once the clustering step achieved, the system will be able to select one sentence in each cluster, dealing efficiently with **centrality** and **diversity**.

Sentence similarity (2)

Similarities between sentences are computed to detect sentences conveying the same opinions.

The similarity measure is based on Jaccard :

$$\frac{\sum (tf.idf(t \in Di, j) * card(t))}{\sum (tf.idf(t \in Di) * card(t)) + \sum (tf.idf(t \in Dj))}$$

We also use WordNet to determine if two terms are equal. In the case they are synonyms, they are weighted by 0,7. The importance of such a use of WordNet has not been studied yet.

Results

Our system is ranked **third** of systems not using snippets, and eleventh of all systems.

Pyramid : 0.169 (best : 0.181, last : 1.101)
Readability : 4.455 (best : 5.318, last : 2.636)

The scores are too erratic. It seems that they depend greatly on summaries length : the summaries we created are, for those worst ranked, very long. A method is needed to better select sentences.

Sentence Selection and Reranking (4)

We have chosen to **extract the sentence** we assume to be **central** in every cluster. We consider **centrality** in a cluster as the ratio between the **similarity to query** and the **distance to the cluster center**.

The sentences from the **most populated clusters** are **ranked first**. Then, we group the sentences following their opiated polarity (cf Michel Génèreux's system) and present the sentences in the queries order.