The Computational Linguistics Summarization Pilot Task: TALN-UPF system description and evaluation

Francesco Ronzano, Beatriz Fisas, Horacio Saggion
NLP Research Group
Universitat Pompeu Fabra
Barcelona, Spain
{francesco.ronzano,beatriz.fisas,horacio.saggion}@upf.edu

Abstract

This document introduces the system built by TALN-UPF team (TALN-UPFsys) to participate in the Computational Linguistic Summarization Pilot Task (CLsumm), organized in the context of the Text Analytics Conference 2014. TALN-UPFsys faces both citation-based document analysis Tasks proposed by CLsumm: reference span identification (Tasks 1a) and discourse facet classification (Task 1b). The system has been trained and evaluated on the corpus of manually annotated documents (CLcorpus) provided by CLsumm organizers. TALN-UPFsys is based on sentence-level computations, thus considering sentences as the main text analysis units. After a brief description of the CLcorpus corpus sanitization and pre-processing steps, we present and evaluate our approach to both CLsumm Tasks.

1 Introduction

The network of article-to-article connections built by citations together with the semantics that characterizes each citation have been leveraged by different scientific text analysis tasks including the assessment of the social relevance of a paper, the targeted selection of contents from scientific articles, single and multi-document summarization.

The Computational Linguistic Summarization Pilot Task (CLsumm), organized in the context of the Text Analytics Conference 2014, proposed two citation-based document analysis Tasks. The first *Task (1A)*, given a citation, required the participants to identify the text spans of the cited paper that are referenced by the same citation; these text spans are also referred to as reference text span. The second *Task (1B)* required to classify each reference text span of the cited paper as belonging to a specific discourse facet, chosen from *Aim*, *Method*, *Implication*, *Results* and *Hypothesis*. A manually annotated corpus, referred to as CLcorpus, was provided by CLsumm organizers in order to train and evaluate the participating systems.

This document describes TALN-UPFsys, the system built by TALN-UPF team to participate in both CLsumm Tasks. In particular, in Section 2 we discuss the process of data sanitization we had to apply to the documents of the CLcorpus so as to enable any further analysis. Section 3 provides an overview of the pre-processing steps we performed to support any more complex computation over the textual contents of each paper. Sections 4 and 5 respectively describe our approach to Tasks 1A and 1B. In these Sections we also provide the results of the evaluation of our system on the CLcorpus. To conclude, Section 6 provides an overview of future venues of improvement.

2 Sanitizing the Computational Linguistic Summarization Corpus

CLsumm organizers provided task participants with the CLcorpus, made of 10 sets of manually annotated papers from the ACL Anthology[1]. Each set includes one cited paper and 5 to 10 citing papers. Each citing paper includes at least one citation to the cited one. The citing papers of CLcorpus in-

[1] http://www.aclweb.org/anthology/

clude 135 citations, with an average of 1,58 citations per citing paper. See Table 1 for more details about CLcorpus.

Docset	Citing papers	Num. cits.	Cits. / paper
C94-2154	5	5	1,000
E03-1020	9	15	1,666
H05-1115	8	12	1,500
H89-2014	8	11	1,375
J00-3003	9	10	1,111
J98-2005	9	21	2,333
N01-1011	8	9	1,125
P98-1081	9	29	3,222
X96-1048	9	12	1,333
-	Avg: 8,4	Sum: 135	Avg: 1,577

Table 1: Number of citing papers and citations of each document set (docset) of CLcorpus.

In citing papers, citations are identified by their citation markers, each one with its formatting style: *(Kogure, 1990)*, *[16]*, *(Gleen el al., 2004)*, etc. For the 135 citations of the CLcorpus, the following set of elements have been annotated by one person:

- the **citation context**, that is the text spans that include the citation marker and usually identify the reason why the paper has been cited;
- one to three **reference text spans** that are the text spans of the cited paper pointing that better reflect the excerpts that are actually referenced by the citation;
- the **discourse facet** characterizing the citation from a predefined set of 5 discourse facets.

For each paper the corpus includes both its original PDF version and the textual output of the PDF-to-text conversion. Stand-off annotations of citation contexts and text spans of the cited papers are provided by means of a CSV file, by specifying the start and end offsets of each text annotation.

When parsing the CLcorpus we experimented several problems that significantly complicated our work:

- **Text encoding**: a small part of the textual documents provided are encoded as UTF-8. Different charset encodings are used including *WINDOWS-1252* and *GB18030*, thus making difficult the implementation of an automated homogeneous textual processing pipeline;
- **Content**: the textual version of the papers, especially with PDF files older than 10 years, presented several text formatting issues: hyphenation problems, words not separated by blank spaces, page headers and footnotes included in the textual flow, etc. The high frequency of these errors prevents analyzing such contents;
- **Stand-off annotations**: in the CSV files, the start and end offsets of each text annotation are not valid offsets of the textual version of the related paper. As a consequence, in order to retrieve the annotated texts, it is necessary to search them manually.

In order to solve all these issues and enable the automated processing of the annotated textual contents of the CLcorpus, we had to perform a heavy sanitization process. In particular we carried out the following steps:

1. conversion of each paper of the corpus from PDF-to-text by means of Poppler[2], a robust PDF-to-text conversor ;
2. manual correction of the PDF-to-text conversion errors in order to get a clean textual version of each paper;
3. by inspecting the textual contents of each CSV files, manual propagation of the annotations of all citing and cited papers to the clean textual version of each paper.

In this way, we generated the sanitized version of the CLcorpus that we used as input for any further textual analysis.

3 Documents pre-processing pipeline

In this Section we provide a brief overview of the pre-processing steps that we perform over each paper of the sanitized version of CLsumm corpus (both citing and cited ones). In this way, we enrich papers with explicit linguistic and semantic metadata that will support the actual text analysis of Task 1A and 1B.

We parse the papers by the following sets of text analysis tools:

[2]http://poppler.freedesktop.org/

1. **Custom rule-based sentence splitter**, to identify candidate sentences that will be validated or rejected by the following pre-processing steps;

2. **Tokenizer** and **POS tagger**. We exploit the ANNIE NLP tools for English, integrated in GATE[3] ;

3. **Sentence sanitizer**, to remove incorrectly annotated sentences, relying on a set of rules and heuristics;

4. **Sentence TF-IDF vector calculator**, useful to associate each sentence with a TF-IDF vector. The IDF values of the terms of each document are computed by considering a corpus including all the papers of the document set the document belongs to (up to 9 citing papers and one reference paper).

4 Task 1A: reference span identification

In Task 1A, starting from a citation, participants have to analyze the cited paper to point out one to three **reference text spans** that identify the excerpts of the cited paper that are actually referenced by that citation.

For each citation we retrieve from CLcorpus the citation context identified by the annotator. Then, we select the sentences of the citing paper that overlap totally or partially the citation context. These sentences are referred to as the citation context sentences (CtxSent1,..., CtxSentN). We associate to each sentence of the cited paper a *score* equal to the sum of the TF-IDF vector cosine similarities computed between that sentence and each sentence belonging to the citation context (CtxSent1,..., CtxSentN). We choose as reference text spans the N sentences of the cited paper with the highest *score*. In the remaining part of this Section we present some experiments to evaluate the performance of our approach when N, the number of cited paper sentences with highest *score* to include in the reference text span, varies.

[3]https://gate.ac.uk/ie/annie.html

CONTENT FROM THE FOLLOWING SECTION NOT YET ADDED TO PROCEEDINGS PAPER

EVALUATION. We evaluate our reference text span identification approach against the gold standard reference text spans manually annotated in CLsumm corpus. To this purpose, we use the F1 score defined by the organizers of the BioMedical Summarization Track of the Text Analysis Conference 2014, since this Track proposed the same Task 1A of CLsumm over a set of papers from the biomedical domain. Such F1 score quantifies the offset-based overlap of pairs of automatically identified and manually selected reference text spans.

In our evaluation we consider four sizes of reference text spans, ranging from the 2 sentences to the 5 sentences of the cited paper with highest similarity to the citation context. The results of our evaluation are summarized in Table 2.

Docset	TOP 2	TOP 3	TOP 4	TOP 5
C90-2039	0,087	0,097	0,153	0,134
C94-2154	0,000	0,096	0,110	0,101
E03-1020	0,087	0.099	0.106	0,093
H05-1115	0,017	0,112	0,106	0,093
H89-2014	0,214	0,196	0,178	0,152
J00-3003	0,121	0,103	0,084	0,072
J98-2005	0,145	0,105	0,083	0,068
N01-1011	0,125	0,107	0,128	0,167
P98-1081	0,104	0,105	0,086	0,072
X96-1048	0,205	0,175	0,153	0,156
AVG:	0,111	0,120	0,121	0,116

Table 2: Variation of the F1 score when the reference text span is identified by considering the 2/3/4/5 sentences of the cited paper with highest similarity to the citation context. F1 score is computed both by document set and averaged across all document sets.

From Table 2 we can notice that our approach obtains the best average result when we select a reference text span that includes the 4 sentences (TOP 4) that are most similar to the citation context. In general, apart from the TOP 2 sentence selection of the document set C94-2154, we can prove that our approach always manages to identify a set of reference text spans that partially overlaps the gold standard ones. In 4 document sets (H05-1115, H89-2014, J00-3003 and X96-1048) we can notice that the best F1 score is obtained by selecting only the TOP 2

sentences. Adding other sentences to the citation context lowers the F1 score. This happens because the 3rd, 4th and 5th most similar sentences are not included in the gold standard reference text spans, thus the precision decreases.

5 Task 1B: discourse facet classification

The purpose of Task 1B is the identification of the **discourse facet** of each reference text span of the cited paper. The discourse facet is the aspect of the cited paper referenced by the citing one. CLcorpus associates to each manually annotated reference text span one among 5 discourse facets: *Aim*, *Method*, *Implication*, *Results* and *Hypothesis*.

We face Task 1B as a sentence classification task. From the CLcorpus, we select the sentences of the cited papers that overlap totally or partially a manually annotated reference text span. We characterize these sentences by the discourse facet that is manually associated to the overlapping reference text span. As a consequence we get a set of 266 cited papers' sentences, each one characterized by a discourse facet (see Table 3).

Docset	Citing papers
Aim	46
Hypothesis	1
Implication	25
Results	29
Method	165
TOTAL:	266

Table 3: Discourse facet of the sentences of cited papers belonging to a manually annotated reference text span.

Considering this dataset we build and compare several sentence classifiers in order to automatically associate to each sentence belonging to a reference text span its discourse facet. Our best sentence classifier obtains an averaged F1 of 0,719. In the rest of this Section we discuss our evaluation of different classification algorithms.

EVALUATION. In order to automatically classify the discourse facet of the sentences belonging to reference text spans, we model each sentence as a word vector that includes lemmas, bigrams and trigrams. When we compute these word vectors we do not remove stopwords.

Once obtained the word vector representation of the sentences, we compare three classification algorithms by a 10-fold cross validation over the set of 266 cited papers' sentences (see Table 3): *Naive Bayes*, *Support Vector Machine* with linear kernel and *Logistic Rgression*. The results of this comparison are shown in Table 4.

Disc. facet	NB	SVM	LR
Aim	0,725	0,734	0,732
Method	0,706	0,826	0,828
Implication	0,049	0,000	0,200
Results	0,509	0,533	0,533
Hypothesis	0,024	0,000	0,000
WEIGHED AVG. F1:	0,623	0,698	0,719

Table 4: Comparison of discourse facet classification algorithms (F1 score): *Naive Bayes* (NB), *Support Vector Machine* with linear kernel (SVM) and *Logistic Rgression* (LR).

CONTENT BEYOND THIS POINT NOT YET ADDED TO PROCEEDINGS PAPER

We can notice that the best performing sentence classifier is the *Logistic Regression* that obtains a weighted average F1 score of 0,719. This classifier obtains good performances for the classes *Aim*, *Method* and *Results*. With respect to these classes, the F1 score for the class *Implication* is instead considerably lower. This probably happens because the examples of *Implication* sentences provided are not enough to properly characterize this kind of class. The F1 score of the class *Hypothesis* is equal to zero, because there is actually only one sentence classified as *Hypothesis* among the 266 sentences of the training set. In general, we believe that the performances of these classifiers can considerably benefit from a greater and more balanced training set.

6 Conclusions and future work

In this document we introduced the system we built to participate to the Computational Linguistic Summarization Pilot Task (CLsumm), organized in the context of the Text Analytics Conference 2014. After some remarks about the CLcorpus sanitation we had to perform, we described the text analysis methodologies we adopted to face the two CLsumm Tasks (Tasks 1A and 1B). We relied on sentence similarity measures to face Task 1A and on a sentence

classifier to deal with Task 1B. We provided an evaluation of both Tasks by comparing the results of our system to the gold standard provided by the CLcorpus. In our future investigation we would like to experiment alternative text similarity approaches to identify reference text spans (Task 1A), taking into account richer structural and contextual information concerning each citation. To improve sentence classification results (Task 1B) we plan to experiment with new feature sets, tailored to the specific task. Moreover, to better validate any approach to these tasks it would be essential to get access to a greater corpus of annotated documents.

Acknowledgments

The research described in this paper is funded by the EU Project Dr. Inventor (FP7-ICT-2013.8ERAGE.1 project number611383), the Project TIN2012-38584-C06-03 of the Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain and the Program Ramón y Cajal 2009 (RYC-2009-04291).