Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Automatic detection of key events from daily news based on recurrent information analysis

MASTER

NATURAL LANGUAGE PROCESSING

Speciality :

Multilingual Engineering

by

Xianfan ZHANG

Thesis Director :

Cyril Grouin

Supervisors :

Amanda Bouffier

Michel Bernardini

Perrine Guy-Duché

2016/2017

Contents

List of figures 5

List of Tables 5

Abstract 7

Acknowledgements 9

Introduction 11

I Context 15

1 Related Work 17
1.1 Introduction . . . 17
1.2 State-of-the-art of event detection . . . 17
1.3 Sentences similarity measures . . . 19
1.4 Conclusion . . . 20

2 Corpus 21
2.1 Annotations and Experiment dataset . . . 21
2.2 Corpus limitations . . . 22

II Experiments 25

3 Methods 27
3.1 Introduction . . . 27
3.2 Sentence-level similarity measure . . . 28
3.3 Document similarity . . . 32
3.4 Headline filter . . . 33
3.5 Conclusion . . . 33

4 Implementation 35
4.1 Introduction . . . 35
4.2 Doublets removing . . . 35
4.3 Data noise reduction . . . 36

4.4 Labeling 37
4.5 Named entity normalization 38

5 Results and discussion **41**
5.1 Introduction 41
5.2 Results 41
5.3 Discussion 43

Conclusion **45**

Bibliography **47**

A Annex **49**
A.1 Benchmark 49

List of figures

0.1 Example of target application 12
2.1 Example of PDF conversion issue 23
3.1 Document clustering schema 28
3.2 Similarity score calculated by equation 3.5 with alpha$\in [0, 1)$ 30
3.3 Logarithmic growth 31
3.4 Similarity score calculated by equation 3.1 with alpha$\in [0, 1)$ 32
4.1 A chart transformed incorrectly to raw text 37
4.2 Labelled document by Analytics2 38
5.1 Key events detected on June 20th, 2017 42
5.2 Key events detected on October 3rd, 2017 43

List of Tables

1.1 A list of sentences from news articles on June 21st, 2017 19
2.1 Composition of corpora 21
3.1 Evaluation at the sentence-level 32
3.2 Evaluation on the news of June 21st, 2017 33
4.1 Regular expression for eliminating bylines 37
5.1 Evaluation on the news of June 20th, 2017 41
5.2 Evaluation on the news of October 3rd, 2017 42
A.1 Event groups of news flow on June 21st, 2017 50
A.2 Event groups of news flow on June 20th, 2017 50
A.3 Event groups of news flow on October 3rd, 2017 51

Abstract

The purpose of this study is to detect the seminal events and relevant documents from daily news. Our approach is based on the hypothesis that if two documents share sentences describing the same facts, they are likely to refer to the same event. Thus, we evaluate the relevance between documents by comparing the textual similarity at the sentence-level. We propose a similarity measure deriving from Jaccard similarity coefficient. The results indicate that this approach is efficient for detecting the "*micro*" events.

Key words: *news monitoring, event detection, sentence similarity measures, event-based clustering, text mining*

Cette étude consiste à détecter les événements majeurs et les documents correspondants dans la presse quotidienne. Notre approche est basée sur l'hypothèse que si deux documents partagent des phrases décrivant les mêmes faits, ils sont susceptibles de faire référence au même événement. Ainsi, nous déterminons la proximité entre les documents en comparant la similarité textuelle au niveau des phrases. Les résultats indiquent que cette approche est efficace pour détecter les "*micro*" événements.

Mots-clefs: *détection de faits d'actualité, détection des événement, similarité entre phrases, clustering, fouille de textes*

Acknowledgements

I would like to express my gratitude to all those who supported and helped me during the internship and the writing of this thesis, in particular my supervisors, Michel Bernardini and Perrine Guy-Duché, as well as all my colleagues in LEONard Team.

Especially, I like to thank Amanda Bouffier, for her insightful criticism and expert guidance which help me overcome the challenging times through the internship.

I would also like to thank my tutor, Cyril Grouin, for his warm-heart encouragement and most valuable advice on the draft of this paper.

Last but not least, I would like to express my thanks to all the professors in INALCO, from whose devoted teaching I have benefited a lot and academically prepared for the thesis.

INTRODUCTION

Context

Digital revolution brings a growing amount of available online information, and thus changes our way of browsing the news. Information management needs to adapt with this evolution. As a Euro-zone leading bank, BNP Paribas needs a competitive intelligence strategy in order to mitigate the risks and to explore external growth opportunities. In this context, the project LEOnard was launched by Michel Bernardini in 2004, aiming at diffusing and sharing information among the collaborators of BNP Paribas. A large amount of news articles are collected in platform from two sources:

— newspapers selected by the company
— web articles crawled by a monitoring tool (KB Crawl)

In order to facilitate the search, all these documents integrated in the platform need to be indexed. Users can thus filter their search by source, language, date or category. All these functionalities involving text mining are carried out through the tools developed by Expert System. Their software suite Luxid is based on the principle of Skill Cartridge, providing different strategies to implement applications according to the needs of companies.

Our study relying on the internship at LEOnard team consists of exploring an application that monitors news articles to detect significant events affecting decision-making process. Some online portals like Wikipedia Current Events Portal take the community efforts to list manually the significant events and relevant news articles. There are also commercial news aggregators like Google News which selects the real time “Top stories” based on their ranking algorithm and users past activities. Our objective aims at implementing a similar application based on textual content of selected newspapers. As shown in Figure 1.1, each event extracted will have a top line as description and related coverage.

It needs to be emphasized that the work presented in this thesis is essentially empirical, and we have built a proof of concept afterwards.

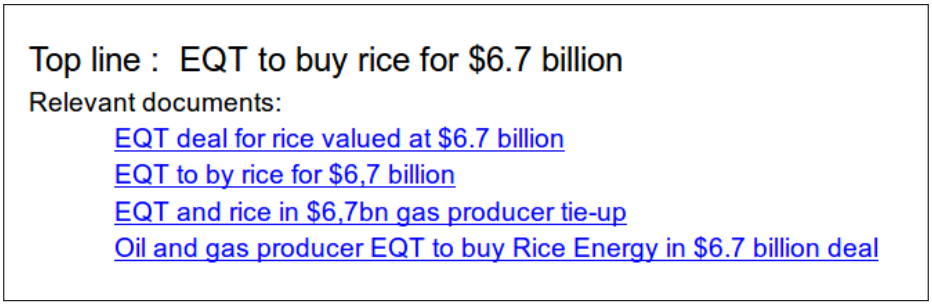


Figure 0.1 – Example of target application

What is a key event?

An event describes usually what happened and corresponds to a change of status [Arnulphy, 2011]. The conventional view of event structure within much of linguistic theory is that a predicate is given a set of arguments and associated diacritics indicating how they are realized [Pustejovsky, 1998]. From the perspective of journalism, a news story about an event needs to convey information such as "where it took place", "who was involved", "when it happened", "how it occurred", "why it came up" and "what are the effects" [Papka, 1999]. On the basis of this point of view, we define an event as a cohesive structural unit containing a set of elements:

Event = {location, participants, action, time, cause, effects }

In economic and political news stories, the participants and the location are mainly represented by named entities. Theses elements are relatively simple to be recognized by automatic processing. The causes and effects are more difficult to be identified. They can be constructed by deverbal nouns, phrases or sentences.

Different individuals will have different *a priori* criteria of what makes an event significant. In general, newspapers are manually curated by editors, the seminal events are recurrent information among articles of the day. Considering that our documents are from reliable sources selected by the company, the importance of an event can rely on the quantity of articles in which they are mentioned. Our main task is to cluster the news documents if they discuss the same event.

Headlines filter

As mentioned in section 1.1, each event will be displayed with a brief description and relevant documents, our second task is selecting one of the articles' headline to represent the event. Some headlines do not convey the essential elements of an event. For example, we found a review article about that Australia suspends Syria air operations, whose headline is "Skirmishing over Syria", with no explicit mention of Australia and air operations. In another example, the headline "Cadmium case

turns toxic for Barclay" referred the fraud charge as "Cadmium case". Users can't clearly figure out what happened with this kind of descriptions." If we consider a set of articles concerning Barclays and fraud charge, we used the headline such as "Barclays and ex-managers accused of crisis-era fraud".

Outline

The rest of this work is organized as follows : chapter 1 gives an overview of the related work on which our method is build; chapter 2 elaborates on the different aspects of the corpus; chapter 3 describes the strategy for solving the problem; the details of the implementation are discussed in chapter 4; chapter 5 is devoted to analyze the results and give a perspective for the future work. Finally, a summary of this study is listed on the conclusion

Part I

Context

CHAPTER 1

RELATED WORK

Contents

1.1 Introduction **17**
1.2 State-of-the-art of event detection **17**
1.2.1 Topic detection 18
1.2.2 Event coreference 18
1.2.3 Problem definition 19
1.3 Sentences similarity measures **19**
1.4 Conclusion **20**

1.1 Introduction

This chapter presents different studies which this work builds upon. In the first section, we figured out the main issue through the discussion of previous researches on event detection. Then we investigate several approaches of sentence similarity measures for solving our issue.

1.2 State-of-the-art of event detection

The problem of event detection is far from being new. This is a classical problem in many contexts. Some of them are carried out in a specific framework, with a number of predefined categories to which the events are assigned. This is the case for example in the ACE program [1], in which event annotations include only those that can be defined under a certain ontological structure. As we focus on the emerging topics in the news of the day, these studies are not discussed in this dissertation. In this section, we present the closest studies, either done within an event detection framework or not, but aiming at clustering the similar facts across documents.

1. https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

1.2.1 Topic detection

Topic detection is a subtask of the Topic Detection and Tracking (TDT) project, consisting of clustering retrospectively the entire corpus. Baseline topics models such as Latent Dirichlet Allocation have served as a very valuable tool for topic detection. Topic model is a type unsupervised learning algorithm capturing hidden semantic structures within text and assigning a document to a mixture of topics. A single document is characterized by several topics of different probabilities. Therefore, topic model can give more realistic results than other clustering algorithms for topic assignment. The study [Dridi and Lapalme, 2013] presents a system for detecting events from Twitter data, where LDA is applied for grouping the similar tweets. Another study [Guan et al., 2016] proposes a extended document clustering which integrates K-means algorithms and LDA modeling into a unified framework to achieve the overall best performance. In these classic document clustering approaches, documents are usually represented as a multiset of words, disregarding the word context, and thus, ignoring the relations of terms. Even though bag-of-ngrams considers the word order in short context, it suffers from the variation and the dispersion of data. This kind of models are efficient to detect the general topic of the texts, but they are insufficient to capture all semantics. In our case, we need to be able to discern accurately the semantic resemblance which is significantly stronger than simple topic similarity.

1.2.2 Event coreference

The event coreference task consists of finding clusters of event mentions that refer to the same event. Two event mentions are coreferential if they have the same event properties and share the same event participants. Different methods are employed to annotate theses properties. An approach is proposed by [Bejan and Harabagiu, 2010], relying on a supervised methods that explore various linguistic features in order to decide if a pair of event mentions is coreferential or not. The linguistic features include lexical features, class features (e.g the Part-Of-Speech), WordNet features (i.e.synonymous relation between terms) and semantic features (i.e. semantic roles). In another study, supervised classifications of event features are proposed for event coreference resolution [Cybulska and Vossen, 2015]. The first step of this approach is filling an event template which contains five slots (i.e. Action, Time, Location, Human Participant, Non-human Participant) for each documents. Subsequently, supervised classifiers determine whether pairs of document templates contain any corefering event mentions. Both of the studies mentioned above gather event information initially at the sentence-level, and then accumulate this information to document-level processing.

1.2.3 Problem definition

As the conventional document clustering algorithms suffer from several shortcomings, we turn to adapt the event coreference resolution approach. The event mentions are collected at the sentence-level where the terms are imposed with syntactic constraints. This method is more fine-grained than working at the document-level.The sentences which describe the events can be considered as a transverse segmentation of the documents content. If a fact is presented in particular sentences in different documents, we expect them to be clustered in the same group. However, determining whether two sentences are coreferential is faced with many challenges. Due to some pragmatic constraints (see Section2.2), we propose an operational and resistant approach for addressing these challenges by comparing the sentence similarity. Table 1.1 shows a list of sentences from different news documents on June 21st about the fraud charge of Barclay's ex-executives. These sentences illustrate that the words used to describe a fact do not vary much among documents describing the same event. In the next section, we present the techniques of sentence similarity measure that can be adapted to identify this kind of information reuse.

DocID	Sentence
34339	The former banker was charged yesterday by the UK Serious Fraud Office with conspiracy to commit fraud over emergency cash injections that saved Barclays at the height of the 2008 financial crisis.
34490	The Serious Fraud Office on Tuesday charged Barclays and some former top executives with conspiracy to commit fraud and unlawful financial assistance.
34436	The Serious Fraud Office charged the bank, the former chief executive, Mr. Varley, former senior investment -bank executive Roger Jenkins and two other former executives with conspiracy to commit fraud.
34309	Barclays, its former chief executive and three other ex-executives have been charged by UK authorities with fraud related to the emergency cash injections that saved the bank from a government bailout at the height of the 2008 financial crisis.

Table 1.1 – A list of sentences from news articles on June 21st, 2017

1.3 Sentences similarity measures

Measuring the similarity between sentences is the basis of most text-related tasks such as question answering application, plagiarism detection, multi-documents summarization. Most of the methods focus on comparing similar parts of two input sentences. The baseline measure (i.e, word overlap measure and the variants) is evaluated in many studies. [Metzler et al., 2006] compares the performance of simple word overlap fraction with the relative-frequency measures and probabilistic models at different similarity levels. The simple word overlap fraction is defined as

the proportion of words that appear in both sentences normalized by the sentence's length. The result shows that this baseline function performs best on the similarity level where the pair of sentences convey some specific facts. Furthermore, on other similarity levels, no other technique was able to significantly outperform the baseline measure. Another study [Adam and Suharjito, 2014] evaluates several variants of word overlap measure. Calculating the overlapping proportion while considering the Part-Of-Speech element of terms gives more stable result than using the other methods.

1.4 Conclusion

In our study, we attempt to cluster the documents on the same event by combining the event coreference approach and sentence similarity measures. Two sentences are considered as relevant if they contain the mentions refer to the same event. While instead of resolving the coreferential event mentions, we turn to detect the similarity between sentences using word overlap measures.

CHAPTER 2

CORPUS

Contents

2.1 Annotations and Experiment dataset **21**

2.2 Corpus limitations **22**

Our experiment corpus is English news flow supplied by LEOnard platform, coming from three sources: Financial Times, Wall Street Journal and The Guardian, which provide over a hundred document converted from PDF format each day. The Wall Street Journal publishes separately the Asian and European edition, where there are same articles share the same content but are edited differently. We explore our method using the articles published on June 21st, 2017, and analyze the global performance by evaluating the flows of three days: June 20th, June 21st and October 3rd.

Corpora	Date	n° of documents	n° of event groups
Experiment corpus	2017-06-21	145	11
Evaluation corpus 1	2017-06-20	102	6
Evaluation corpus 2	2017-10-03	105	6

Table 2.1 – Composition of corpora

2.1 Annotations and Experiment dataset

For each flow, we distribute manually news articles into event groups, and there are around ten groups emerging from the flow of the day. Each of them contains at least two documents that refer to the same event. These benchmark groups are found in Annex, section A.1.

As the first stage of our method lies on the sentence level, we prepare a dateset of 150 sentence pairs which correspond to two categories :

1 : Sentences from documents concerning the same event

s1: **The former banker was charged yesterday by the UK Serious Fraud Office with conspiracy to commit fraud over emergency cash injections that saved Barclays at the height of the 2008 financial crisis.**

s2: **The Serious Fraud Office on Tuesday charged Barclays and some former top executives with conspiracy to commit fraud and unlawful financial assistance.**

2 : Sentences from documents concerning different events but they share some words or named entities.

s3: **Signalling a desire for more European unity ahead of an EU summit this week, Ms Merkel also warned the UK that her priority in Brexit talks was to avoid splits among the blocs remaining 27 member states.**

s4: **German Chancellor Angela Merkel for the first time sketched out the outlines of a bargain with France on fixing the governance of European single currency, in the clearest sign yet that the two biggest eurozone countries are inching toward reconciling sharply different views on the matter.**

2.2 Corpus limitations

This corpus itself has some limitations, which create hurdles of several kinds. News reporting follows the principle of language economy, meaning that some information previously communicated within a unit of discourse, will not be mentioned again, unless pragmatically required. As a result, biases could not be ignored and limitations of corpus affect the choice of methodology.

The first limitation is that the corpus has little redundancy. Among events extracted from the flow each day, most of them are referred only in one or two documents. Even the mostly mentioned events have merely on average five relevant documents. For example, the mostly mentioned event on June 20th is that Brexit discussion begins, with merely three relevant documents. As a result, it is arduous to resort to statistical methods.

The second limitation is the data noise caused by PDF converter. Firstly, since the headings are a special kind of text and are not as rigidly governed by conventions of punctuation, sometimes they do not end with a full stop. Considering that our corpus is plain text in format TXT that ignores the content structure, the section headings and text body will be concatenated without boundaries. Furthermore, the journal may contain complex layout which is not preserved by raw text. As a result, some paragraphs are misplaced when converting PDF to TXT. An example of this formatting issue is shown in Figure 2.1. The headline, the subhead and authors' in-

formation are attached to the lead paragraph. The sentence marked by green color is split by a photo caption and another paragraph. The examples mentioned above are not exhaustive. A large variety of conversion issue makes it difficult to get rid of all the noise by automatic processing. In consequence, some sentences are not correctly segmented especially those which convey the essential information (i.g.section headings and the beginning of the body). The errors of sentence segmentation will distort the result reliability of Part-Of-Speech Tagging and semantic role labelling. For this reason, we road to a lenient approach instead of analyzing the semantic roles of participants or the syntactic structure of events.

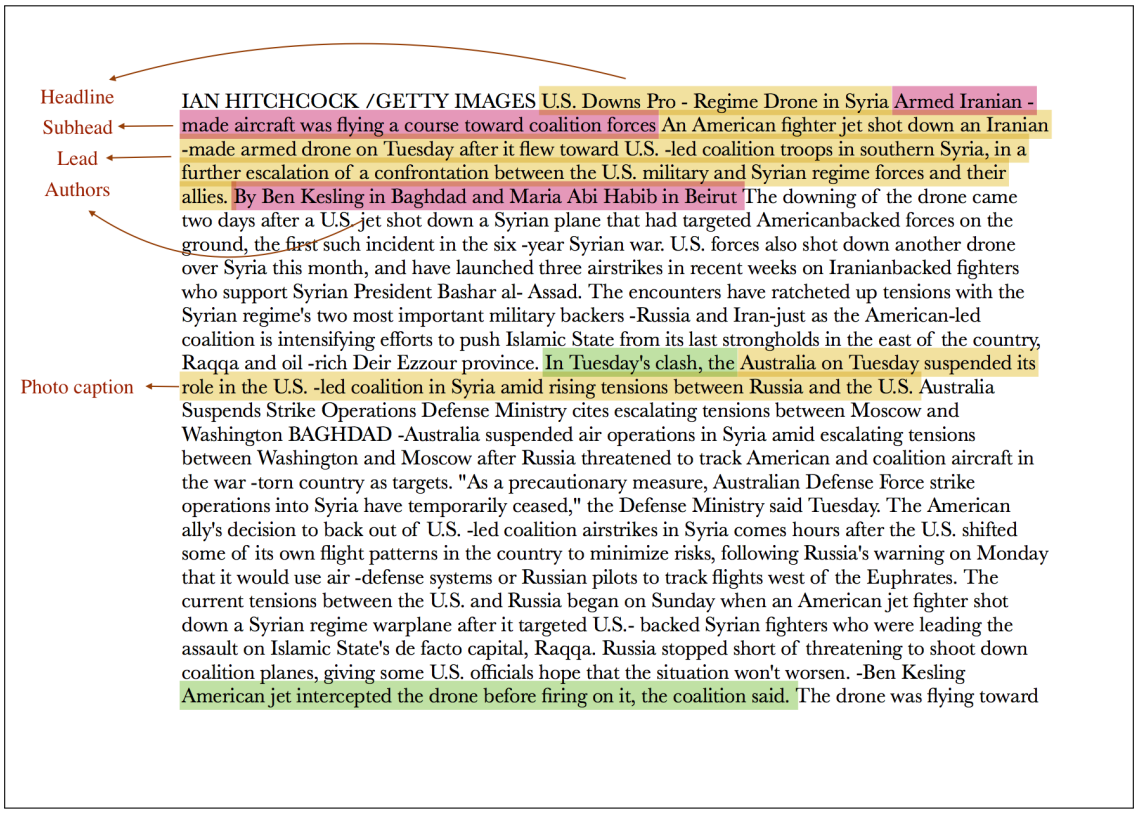


Figure 2.1 – Example of PDF conversion issue

Part II

Experiments

CHAPTER 3

METHODS

Contents

3.1 Introduction **27**

3.2 Sentence-level similarity measure **28**

3.2.1 Relative contribution of named entity and common word . . 28

3.2.2 Probability of information redundancy 30

3.3 Document similarity **32**

3.4 Headline filter **33**

3.5 Conclusion **33**

3.1 Introduction

In this chapter, we mainly present our event-based clustering strategy. Our method consists of two stages. At the first stage, we calculate the similarity of sentence pairs across documents. If the similarity exceeds a certain point ($threshold_s$), we consider that two sentences concern the same event and they are similar. The next stage is identifying if two documents share factual content referring to the same event. For each pair of documents, we calculate the similarity score of documents by combining the similarity score of similar sentences (the outcome of the previous stage). Then two documents will be assigned to the same event group if their document similarity score is greater than a threshold ($threshold_d$). The procedure is manifested in Figure 3.1.

Firstly, we elaborate on the methods and the parameters used to measuring the similarity between sentences. Secondly we attempt to capture numerically the extent to which documents covering the same event. We evaluate the configurations by calculating precision and recall. In our case, precision is the most important index for measuring the effectiveness of the methods. Lastly, we carry out a linguistic analysis on the headlines in order to choose those that encapsulate the news content.

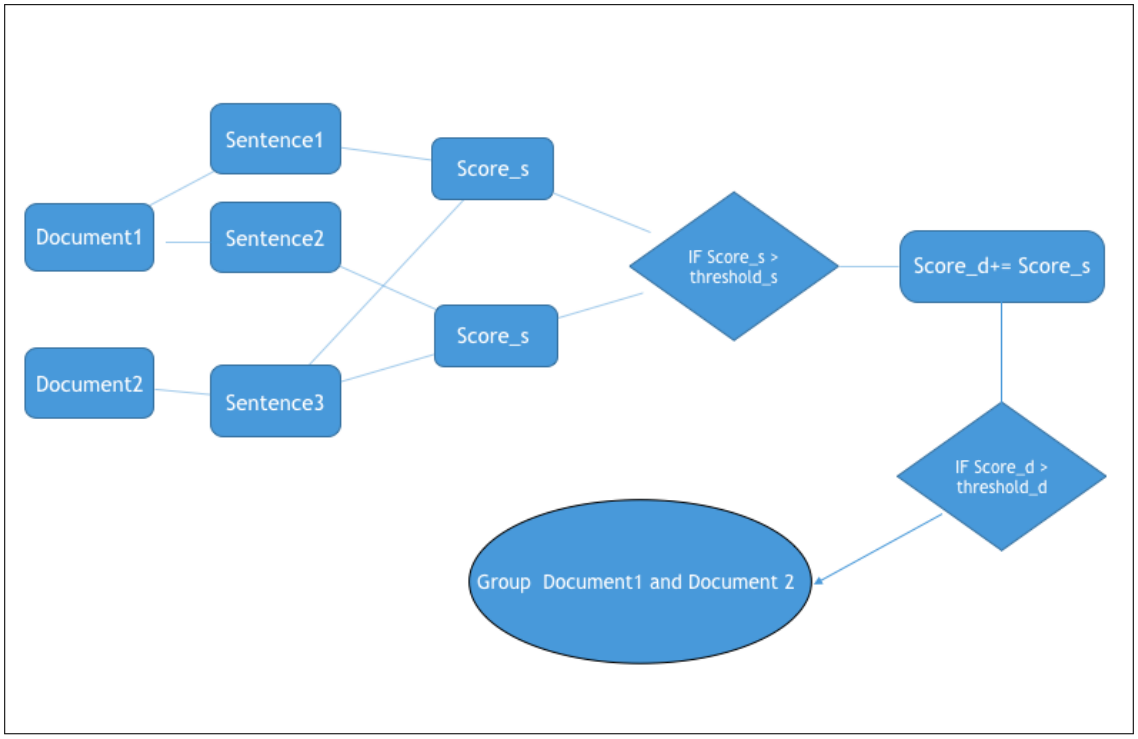


Figure 3.1 – Document clustering schema

3.2 Sentence-level similarity measure

We devise an equation in order to calculate the similarity score S(s1,s2) between two sentences, s1 and s2, in different documents:

$$S(s1, s2) = \alpha \log_2(|ne_{s1} \cap ne_{s2}|)Sim_{ne} + (1 - \alpha) \ln(|w_{s1} \cap w_{s2}|)Sim_w \tag{3.1}$$

where $|ne_{s1} \cap ne_{s2}|$ is the number of the shared named entities in the pair; Sim_{ne} is the similarity score of named entities;$|w_{s1} \cap w_{s2}|$ is the number of the shared common words in the pair; Sim_w is the similarity score of common words; α is a coefficient for balancing the weight of named entities and common words in similarity measuring. In following subsections, we present the assumptions based on which the equation is built.

3.2.1 Relative contribution of named entity and common word

Our event similarity measure is a variant of word overlap measures, which are the baseline measures that compute similarity score relying on a number of words shared by two sentences. The basic fraction is Jaccard similarity coefficient which defines the similarity score Sim(s1,s2) as the size of the intersection divided by the

size of the union of the words in the two sentences s1 and s2[Priya and M.E., 2013]:

$$Sim(s1, s2) = \frac{|words_{s1} \cap words_{s2}|}{|words_{s1} \cup words_{s2}|} \tag{3.2}$$

We adapt the formulae to our task and calculate separately **Named entities similarity** (Sim_{ne}) and **Common words similarity** (Sim_w) for each sentence pair.

$$Sim_{ne}(\text{s1,s2}) = \frac{|ne_{s1} \cap ne_{s2}|}{|ne_{s1} \cup ne_{s2}|} \tag{3.3}$$

$$Sim_{w}(\text{s1,s2}) = \frac{|w_{s1} \cap w_{s2}|}{|w_{s1} \cup w_{s2}|} \tag{3.4}$$

The hypothesis is that the semantic content carried by named entities is more important for event identification. For example, s1 and s2 use the same words to describe business acquisition. However, they are considered as different event mentions because the doers of the action are different entities.

s1: **Apple has acquired a French startup.**

s2: **Orange has acquired a French startup.**

In natural language, there are opposite examples as presented in s3 and s4, where two sentences describe different events while concerning the same entites. But this case is presumed to be rare in our corpus, because we work on the information reported within one day, where different events involved the same participants rarely coincide.

s3: **Lily got married with Tony.**

s4: **Lily divorced with Tony.**

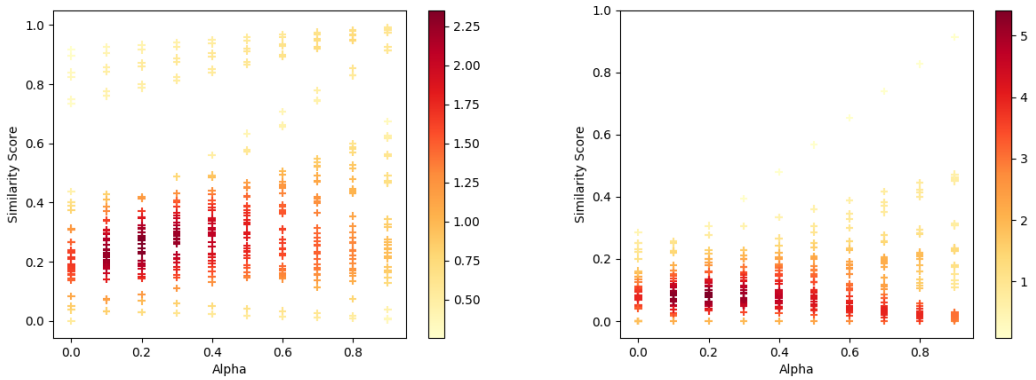
In order to discriminate the contribution of named entities and common words, we define the word overlap measure as a linear combination of **Named entities similarity** (equation 3.3) and **Common words similarity** (equation 3.4). The relative contribution of named entities and common words measures is controlled by a coefficient alpha (see equation 3.5). The *common words* include terms with POS tags as follows:

— /Term/COMMON-NOUN
— /Term/ADJ
— /Term/NOUN
— /Term/VERB
— /Term/Money

$$Sim(s1, s2) = \alpha Sim_{ne}(s1, s2) + (1 - \alpha) Sim_{w}(s1, s2) \tag{3.5}$$

In an attempt to visualize how the coefficient alpha impact the similarity score, we compute the similarity score with alpha ranging from 0 to 1. Then two scatter plot (Figure 3.2a and Figure 3.2b) are constructed from the results with color indicating data density.

Figure 3.1a and Figure 3.1b illustrate that, with the same alpha, the similarity score of sentence pairs describing the same fact are generally higher than those describing different facts. Besides, the alpha is in direct proportion with the score of pairs covering the same fact. While for the pairs covering the different facts, most of the scores decrease with the increase of alpha. The trends confirm our hypothesis that when named entity similarity is weighted more than common word similarity, the score is more discriminating across the similarity levels.



(a) Pair of sentences describing the same event (b) Pair of sentences describing different events

Figure 3.2 – Similarity score calculated by equation 3.5 with alpha$\in [0, 1)$

3.2.2 Probability of information redundancy

However, as seen in Figure 3.1b, there is a small part of sentence pairs describing different events whose similarity score increases with the increase of alpha. These are the pairs containing only one named entity in each sentence, and this named entity are shared in two sentences. As their intersection and union sizes of named entities are equivalent, they will be evaluated as highly similar according to Jaccard fraction. For example, the following sentences (s1, s2) discuss different topics while their named entity similarity will be 1 according the previous formulae.

s1 **If the trend continues, it could shorten the amount of time the government has before it runs out of cash to pay its bills, unless Congress raises the federal borrowing limit.**

s2 **A draft bill, coming before Congress in September, targets crowdfunding, cryptocurrencies and payment technology.**

Suppose that there are two sentence pairs of which the union sizes of words are unknown. The pairs with a large set of shared word have more chance of covering the same facts than the others. It means that the size of intersection affects the probability of information redundancy. We propose to model this effect by the logarithmic growth, and the probability should be considered as a weight to regulate the similarity. This assumption comes from the intuition that, along with the increase of intersection size, the probability of information redundancy has a period of rapid increase followed by a period where the growth slows.

The formulae is extended as shown in equation 3.1, where the named entities similarity is regulated by the binary logarithm ($\log_2(|ne \cap ne|)Sim_{ne}$), and the common words similarity by the natural logarithm ($\ln(|w \cap w|)Sim_w$).

— If two sentences have only one named entity shared, their named entity similarity will be 0 ($\log_2 1 = 0$);

— If two sentences have two named entities shared, the named entity similarity will not be affected by the additional weight ($\log_2 2 = 1$);

— If two sentences have two common words shared, the common word similarity will be lowered ($0 < \ln 2 < 1$).

Figure 3.3 manifests the growth rate of the weighting for named entities similarity (weight1) and the weighting for common words similarity (weight2).

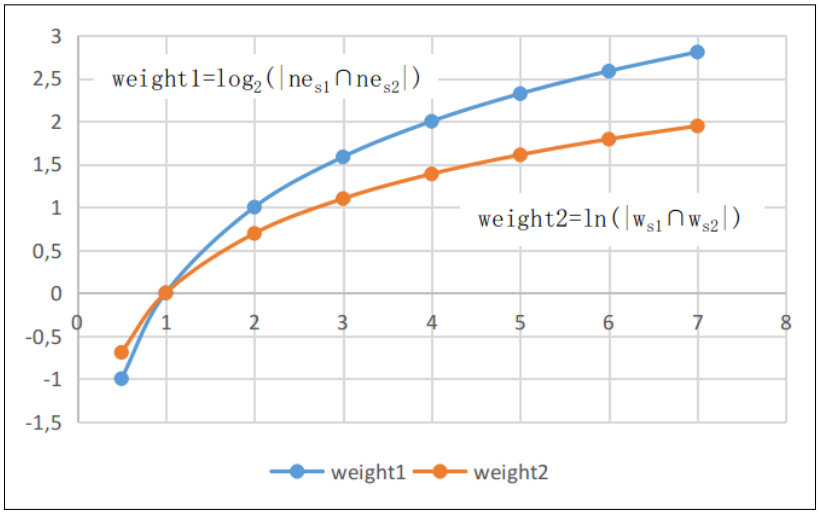
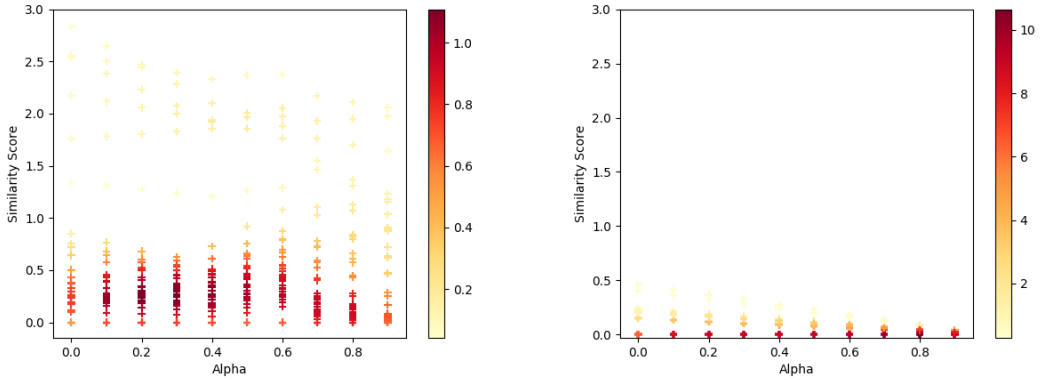


Figure 3.3 – Logarithmic growth

We recompute the similarity scores by equation 3.1 with alpha ranging from 0 to 1 and plot the results in two charts shown in Figure 3.4. The results reveal a definite improvement on drawing a boundary of different similarity levels.

In order to find out the best-performing combination of alpha value and score



(a) Pair of sentences describing the same event (b) Pair of sentences describing different events

Figure 3.4 – Similarity score calculated by equation 3.1 with alpha$\in [0, 1)$

threshold, we calculate the *precision* and *recall* with alpha with range 0 to 1 across the score threshold. For this analysis, we chose to show the best results within the alpha range in Table 3.1. The result illustrates that the method performs the best when named entity similarity is weighted more than other common words similarity with α around 0.6 and $threshold_s$ around 0.2.

α	$threshold_s$	Precision	Recall	F-score
0.3	0.2	0.91	0.70	0.79
0.4	0.2	0.90	0.66	0.76
0.5	0.2	0.92	0.81	0.86
0.6	**0.2**	**0.93**	0.89	**0.91**
0.6	0.3	0.93	0.84	0.88
0.7	**0.2**	**0.93**	0.89	**0.91**
0.8	0.1	1	0.61	0.76

Table 3.1 – Evaluation at the sentence-level

3.3 Document similarity

In this section, we perform a preliminary evaluation on document clustering results by testing different document similarity thresholds $threshold_d$. The key hypothesis in this stage is that two documents referring to the same event are expected to contain pairs of sentences concerning the same event. We suggest building document scores by combining individual sentence-to-sentence scores which exceed the sentence similarity threshold $threshold_s$.

In the previous section, we empirically obtain the configuration ($\alpha = 0.6$, $threshold_s = 0.2$) which produce the best precision and overall score. We use this configuration to select the similar sentences across the documents of experiment cor-

pus. The document similarity score are the total of the similar sentence score. As seen in Table 3.2, the best result is obtained if we group the documents whose similarity score is greater than 1.5. In chapter 5, we carry out this evaluation on two other corpus in Chapter 5, with analysis further detailed.

$threshold_d$	Precision	Recall	F-score
1	0.81	0.86	0.83
1.25	0.81	0.86	0.83
1.5	**0.94**	0.81	**0.87**
1.75	0.94	0.74	0.83

Table 3.2 – Evaluation on the news of June 21st, 2017

3.4 Headline filter

In this section we discuss the strategy to select an appropriate headline for displaying on the platform. Generally, the headline is a brief sentence indicating the nature of article or news story below it, sometimes with auxiliary verbs and articles removed. However, not all the headlines summarize the essential information of an event. In our case, a descriptive headline should be a declarative sentence indicating at least the participants and the action of an event. Thus, syntactically we propose to eliminate:

— headlines that begin with an interrogative pronoun; (e.g. *Why the country is ripe for a start-up boom How iPhone Decade Reshaped Apple*)
— headlines that omit the subject; (e.g. *Hold banks to account*)
— headlines without verb; (e.g. *Skirmishing over Syria*)

Besides, some headlines imply analogy or coreference so that the facts are not revealed. Therefore, we suggest an elimination on the semantic level:

— Eliminate the headlines without an named entities; (e.g. *The former executives caught up in investigation*)
— Calculate the similarity score using the sentence similarity measure; Choose an alternative from pair with the highest similarity score.

3.5 Conclusion

In this chapter, we develop the strategy for detecting the mostly mentioned event from the news of a day. The document clustering is performed from the sentence level to the document level. By testing the experiment corpus, we empirically obtain a favourable configuration ($\alpha = 0.6$, $threshold_s = 0.3$, $threshold_d = 1.5$). Finally for

each group emerged from the clustering, we choose descriptive headline according to a set of syntactic and semantic constraints.

CHAPTER 4

IMPLEMENTATION

Contents

4.1 Introduction 35
4.2 Doublets removing 35
4.3 Data noise reduction 36
4.4 Labeling 37
4.5 Named entity normalization 38

4.1 Introduction

This chapter describes the details of the implementation and the tools we used to realize the project. We focus on the dataset preparation which include, but not limited to the general pre-processing blocks (i.e. tokenization, sentence segmentation, lemmatization, Part-Of-Speech tagging, named entity extraction, stop-words removal). This step is indispensable for improving the performance of our methods.

It is worth to emphasize that the study presented in this thesis intend to develop a strategy which can be implemented by Expert System at a later stage. Therefore, we use preferentially their existing tools for develop the processing schema. The problem is that their Skill Cartridges are folded and some processing procedures could not be concatenated without the extension. For this reason, some tasks are undertaken individually using an open source tool. For example, selecting sentences containing named entities and sentences labelling, as discussed in section4.4, are incorporated in one processing block in general. In our study, they are resolved separately using a tool open sources and a tool supplied by Expert System.

4.2 Doublets removing

In two editions of Wall Street Journal, there are a slice of doublet documents of several types which are expected to be removed:

— Truncated texts: a document is a truncated version of another

— Quasi-doublets: two documents are generally identical except for some minor edits
— Full-doublets: two documents are completely identical

Our sentence similarity measure (mentioned in chapter 3) could not be used to identify the doublets because it ignores the words order. While two articles are considered identical only if they use the same words within the same order. Therefore, we propose to detect the doublets by comparing the bigram overlap. The fraction used to calculate the doublet score(Doublet(D1,D2)) is defined as follows:

$$\text{Doublet(D1,D2)} = \frac{\text{intersection(bigramD1,bigramD2)}}{\text{min(bigramD1,bigramD2)}} \tag{4.1}$$

It is empirically proved that the measure performs the best when the threshold of score are set at 0.8. If the pair of documents obtain a doublet score greater than 0.8, the shorter version is removed.

4.3 Data noise reduction

Since most of the NLP tools are made to handle error-free texts, cleaning the corpus is a indispensable procedure before feeding the corpus files for experimentation. We take note of two typical types of noise, and replace the strings through regular expression using Python re module [1]

Almost all the documents have a short phrase that indicates the name of the author and the place of reporting (i.e. a byline). It is important to remove this information because they can be recognized as participants and location of an event. In most cases, it is prefaced by the wording indicating that this piece of information is the name of the author (e.g."by"). Or else it is written uppercase, while other named entities of the type "person" are in capitals. Four regular expression patterns are applied fetching 149 bylines from the flow of June 21st. Table 4.1 illustrates the patterns with a matching example.

Another type of noise is caused by the numeric data of the charts which is compiled during conversion. For example, The chart presented in Figure 4.1 is transformed into a string "*Potential full inclusion MSCI country weights (%) Russia 3.0 Others Mexico 3.5 India 5.2 Brazil 5.5 S Africa 5.6 Taiwan Source: MSCI*". 89 sequences of noise in experiment corpus are wiped out through RE patterns as follows:

Pattern 1. `((±\.?\d*\s){2,})`

Pattern 2. `(([A-Z]\w*\s){1,2}\d+\.?\d*\s){2,})`

1. The documentation of Python3 re module: https://docs.python.org/3/library/re.html

RegExp	Example
((((b|B)(Y|y))\s)?[A-Z]{3,}((\s|\-|\/)+([A-Z]\.)?[A-Z]{3,})+)	ANNE -SYLVAINE CHASSANY - PARIS
((A|a)dditional\sreporting\sby\s([A-Z]\w+\s){2,4})	Additional reporting by Martin Arnold Lex
(([A-Z]\sw+\s){2,}in\s[A-Z]\w+\scontributed\sto\sthis\sarticle)	Lilian Lin in Beijing contributed to this article
((b|B)y\s([A-Z]\w+\s){2,}in\s([A-Z]\w+)(\sand\s([A-Z]\w+\s){2,}in\s([A-Z]\w+))?)	By Ben Kesling in Baghdad and Maria Abi Habib in Beirut

Table 4.1 – Regular expression for eliminating bylines

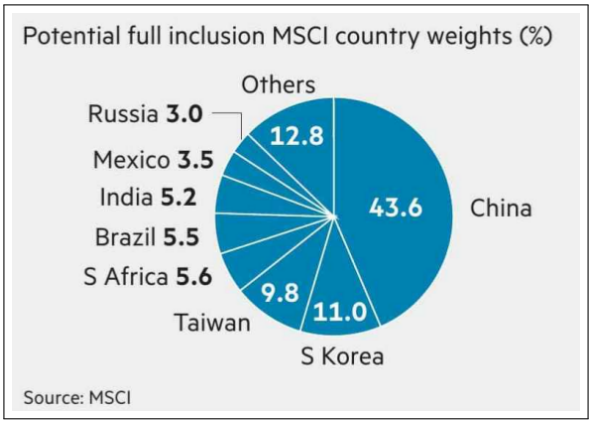


Figure 4.1 – A chart transformed incorrectly to raw text

4.4 Labeling

As discussed in introduction, the participants and the location of an event are always in form of named entity in political and economic spheres. It is not worthwhile to process the textual units without these essential elements. Therefore, we select the sentences which contain at least one named entity of type "person","organization" and "location". Then the event detection turns to all pair combinations of these sentences from different documents. The sentence filter is implemented by a Java program, where we integrate the Stanford CoreNLP toolkit [2].

In the next step, the target sentences are labelled using a Skill Cartridge [3]: Analytics2, which is developed by Expert System [4]. This Cartridge involves format conversion (i.e. output file in TMX format) and morphological analysis. The latter is

2. Stanford CoreNLP : https://stanfordnlp.github.io/CoreNLP/
3. Skill Cartridge is a block of text analytics tool designed to be integrated into different workflows.
4. Expert System is a semantic intelligence company that creates artificial intelligence, cognitive computing and semantic technology software.

carried out by the morphological analyzer, which use the technology XFST (Xerox Finite State Technology). Here are the most important functions serving our experiment :

— tokenization and lemmatization
— removing functional word
— morphological analysis
— morph-syntactic disambiguation

The output of labelling is shown in Figure 3.3.

```
<annotations>
  <annotation name="juice" type="/Term/COMMON-NOUN" s="106" e="111"/>
  <annotation name="producer" type="/Term/COMMON-NOUN" s="113" e="122"/>
  <annotation name="Huiyuan" type="/Term/PROPER-NAME" s="131" e="138"/>
  <annotation name="multinational" type="/Term/COMMON-NOUN" s="140" e="154"/>
  <annotation name="slow" type="/Term/ADJ" s="165" e="169"/>
  <annotation name="adapt" type="/Term/VERB" s="173" e="178"/>
  <annotation name="speed" type="/Term/COMMON-NOUN" s="186" e="191"/>
  <annotation name="market" type="/Term/COMMON-NOUN" s="205" e="211"/>
  <annotation name="change" type="/Term/VERB" s="215" e="223"/>
  <annotation name="Nongfu" type="/Term/PROPER-NAME" s="234" e="240"/>
  <annotation name="score" type="/Term/VERB" s="246" e="252"/>
  <annotation name="expensive" type="/Term/ADJ" s="258" e="267"/>
  <annotation name="fromconcentrate product" type="/Term/NP" s="272" e="296"/>
  <annotation name="fromconcentrate" type="/Term/COMMON-NOUN" s="272" e="287"/>
  <annotation name="product" type="/Term/COMMON-NOUN" s="288" e="296"/>
  <annotation name="package" type="/Term/VERB" s="297" e="305"/>
  <annotation name="health drink" type="/Term/NP" s="309" e="322"/>
  <annotation name="health" type="/Term/COMMON-NOUN" s="309" e="315"/>
  <annotation name="drink" type="/Term/COMMON-NOUN" s="316" e="322"/>
```

Figure 4.2 – Labelled document by Analytics2

4.5 Named entity normalization

The labelling result of named entity can not be use directly, due to their synonymy and ambiguity across documents. For example, the strings U.S. USA and America can all be used to refer to the concept United States of America. Similarly, the string Washington can be used to refer to different entities in our corpus. (e.g.Washington or USA). One approach to addressing these problems is Named Entity Normalization (NEN), which goes beyond the NER task: names are not only identified, but also normalized to the concepts they refer to. Expert System supplies a Skill Cartridge TM360 extracting and normalizing more than 20 types of shared entities with a satisfying performance [5], which can be integrated to implementation of the company. However, this Cartridge could not be merged with the Analytics2 (i.e. the labelling Cartridge) without the customization for the moment. In order to improve efficiently the practicability of our strategy, we do a simple normalization by a Python containing two sections.

The first section is unifying country names by means of a country name list [6].

5. We evaluated the performance of Skill Cartridge TM360 on short text (summaries of articles) in another project through internship, and obtained F-score as 93%

6. List found in git-hub: https://github.com/sshaw/normalize_country/blob/master/lib/normalize_country/countries/en.yml

where each country consists of their aliases, short name and official name. All the alternative versions are transformed into short name.

The second section is identifying the use of **synecdoche** in named entities. The synecdoche is a figure of speech in which a term for a part of something refers to the whole of something or vice versa. The most common synecdoche in news reporting is that the national capital is often used to represent the government or monarchy of a country[Zelizer, 1990]. Therefore, we transform the capital city name to country name if the former do not preceded by the preposition indicating the position. *E.g. Increasingly tense conflicts in the South China and East China seas are entwined with ~~Beijing~~* ***China****'s boarder goal to curb ~~Washington~~* ***America****'s long military supremacy in the Pacific.* Furthermore, another common narrative strategy in news writing is that the government buildings are associated with their occupants. We did not deal with this disambiguation because the buildings can refer to different entities according to the context. E.g., "The White House" can either refer to the President of America or his staff.

CHAPTER 5

RESULTS AND DISCUSSION

Contents

5.1 Introduction **41**
5.2 Results **41**
5.3 Discussion **43**

5.1 Introduction

Our goal in this study has been to list automatically the mostly mentioned events with a brief description and the relevant documents. In chapter 3, we have already evaluated the method performance on experiment corpus at the sentence-level and the document-level, The result is favorable at the sentence-level when the $threshold_s$ is set to 0.2. In this chapter, we firstly evaluate our document clustering performance with two other news flows, using the sentence-level threshold 0.2. In what follows, we analyze the result by comparing the clustered event groups with the benchmark. 3.4. Lastly, we take a global view of this study and discuss what issues could arise in future work.

5.2 Results

The results from the evaluation on news documents of June 20th and October 3rd are summarized in Table 5.1 and 5.2. Same as the result of experiment corpus, the optimized precision values are obtained when the $threshold_d$ is set to 1.5 or higher. However, the overall recall is much more lower than that of the experiment results.

$threshold_d$	Precision	Recall	F-score
1	0.69	0.64	0.66
1.25	0.82	0.64	0.72
1.5	**0.9**	0.64	0.75
1.75	**0.9**	0.64	0.75

Table 5.1 – Evaluation on the news of June 20th, 2017

$threshold_d$	Precision	Recall	F-score
1	0.75	0.88	0.81
1.25	0.9	0.53	0.67
1.5	0.9	0.53	0.67
1.75	0.9	0.53	0.67

Table 5.2 – Evaluation on the news of October 3rd, 2017

Recall that we define an event as a set of elements containing the participants, location, time, as well as the its cause and effects which are sometimes other events. Therefore, we expect to group the documents referring to a seminal event, and those covering all directly related events and activities. In news articles, there are some "*macro*" events which consist of a series of activities. For example, the Catalan independence movement which is widely discussed in the news on October 3rd containing a set of sub-events can be considered as "*macro*". By contrast, the dimission of Uber's executive is relatively "*micro*" because the actants and the action are more specific.

The automatically generated groups are shown in Figure 5.1 and 5.2. Comparing with the manually listed groups (see section A.1), we found that the our method performs unsatisfactorily on "*macro*" events, of which different aspects are usually reported the in different articles. As a result, these articles probably do not have enough term co-occurrence. Turning again to the example of Catalan independence movement, the relevant documents have not been clustered because this news is reported from different perspectives: the current situation in Catalonia and in Spain; the reaction from the European Commission; The position Catalan President, etc. Conversely, the clustering results for the "*micro*" events such as "EQT's acquisition of Rice Energy", "the fraud charge of Barclay's ex-executives" and "the dimission of Uber's executive" are satisfying.

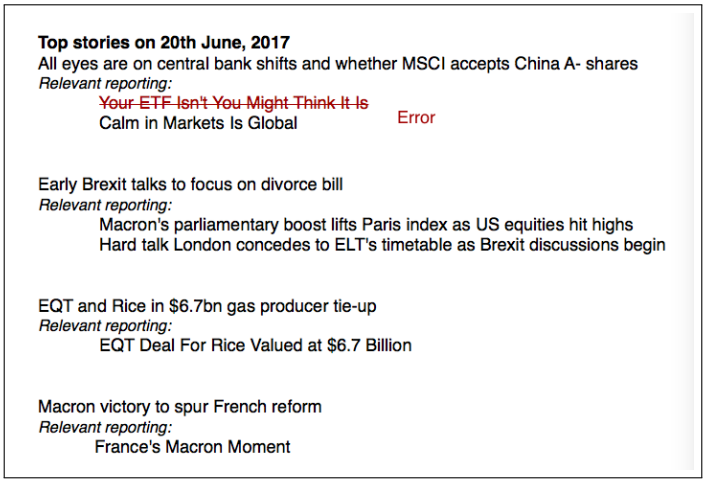


Figure 5.1 – Key events detected on June 20th, 2017

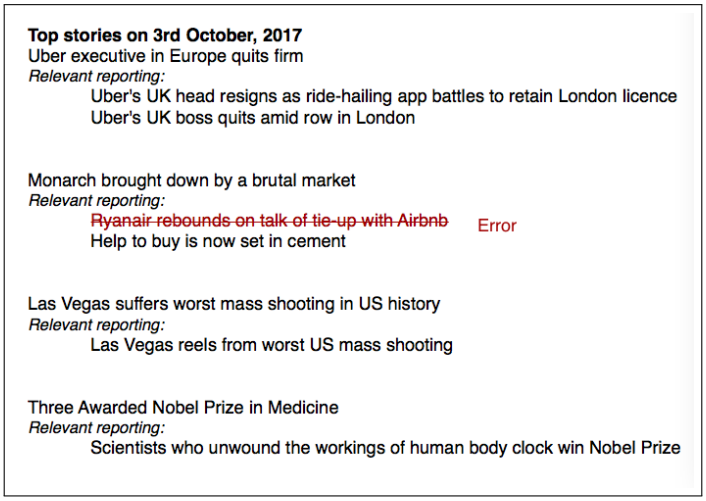


Figure 5.2 – Key events detected on October 3rd, 2017

5.3 Discussion

This study was successful in terms of clustering the documents with a high accuracy. However, not all the seminal events can emerge from the detection because the cluster size of "*macro*" events is sometimes small while only the events with a large document cluster size could be considered as key events. In order to improve the performance of the target application, we need to increase recall or adjust the criteria for an event to be considered important.

Many study indicate that taking account of semantic information implied in the sentences can improve the recall especially in the case of plagiarism detection [Priya and M.E., 2013],[Achananuparp et al., 2008],[Adam and Suharjito, 2014] . The semantic similarity of two sentences is calculated using distance between the terms from a structured lexical database. This approach enables the method to model human common sense knowledge. However, it may reduce the accuracy if a larger and general semantic nets is applied (e.g. WordNet) [Li et al., 2006]. For example, in WordNet, the minimum path length from boy to animal is 4, less than from boy to teacher which is 6, while intuitively, boy is more similar to teacher than to animal. As in our task, the precision is the priory index for evaluating the effectiveness, we did not adapt this approach to our method. In future word, we could perhaps address this weakness by selecting a certain part of hierarchy in the semantic nets to calculate. For example, the words at upper layers of hierarchy with more general semantics will not be taken into account, while the distance between the words at lower layers is more important because they have more concrete semantics.

Conclusion

In this study we explored a method for detecting the significant events from daily news relying on the quantity of relevant documents. We mainly focused on clustering the relevant documents by comparing the textual similarity at the sentence-level. We discussed as well the limitations of the materials and proposed a processing sequence in order to optimize the performance of implementation.

The similarity measure used in this study is a variant of Jaccard similarity coefficient. We devised an equation which distinguishes the contribution of named entities and common words by a coefficient alpha and the logarithm functions with different bases. The document similarity is calculated by totaling the score of similar sentences. The results are controlled by three parameters: a coefficient α balancing the relative contribution of named entities and common words; a sentence-level $threshold_s$ drawing a boundary between sentences of different similarity levels; a document-level $threshold_d$ according to which two documents are considered as relevant.

The best-performing combination of the paramaters is empirically obtained by testing on the English news flow of three days ($\alpha = 0.6$, $threshold_s = 0.2$, $threshold_d = 1.5$). The evaluation shows that the overall precision is favorable, while the recall remains to be improved. The method fails to cluster the documents which refer to the events containing a large set of aspects. Further study may refine the method by calculating the semantic similarity instead of the simple overlapping proportion.

BIBLIOGRAPHY

[Achananuparp et al., 2008] Achananuparp, P., Hu, X., and Xiajiong, S. (2008). The evaluation of sentence similarity measures. – Cité page 43.

[Adam and Suharjito, 2014] Adam, A. R. and Suharjito (2014). Plagiarism detection algorithm using natural language processing based on grammar analyzing. – Cité pages 20 et 43.

[Arnulphy, 2011] Arnulphy, B. (2011). A weighted lexicon of french event names. – Cité page 12.

[Bejan and Harabagiu, 2010] Bejan, C. A. and Harabagiu, S. (2010). Unsupervised event coreference resolution with rich linguistic features. – Cité page 18.

[Cybulska and Vossen, 2015] Cybulska, A. and Vossen, P. (2015). "bag of events" approach to event coreference resolution. supervised classification of event templates. – Cité page 18.

[Dridi and Lapalme, 2013] Dridi, H. E. and Lapalme, G. (2013). Détection d'évènements à partir de twitter. – Cité page 18.

[Guan et al., 2016] Guan, P., Chen, Y. W. B., and Fu, Z. (2016). An improved ant algorithm with lda-based representation for text document clustering. – Cité page 18.

[Li et al., 2006] Li, Y., McLean, D., Bandar, Z. A., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. – Cité page 43.

[Metzler et al., 2006] Metzler, D., Bernstein, Y., Croft, W., Moffat, A., and Zobel, J. (2006). Similarity measures for tracking information flow. – Cité page 19.

[Papka, 1999] Papka, R. (1999). On-line new event detetion, clustering and traking. – Cité page 12.

[Priya and M.E., 2013] Priya, M. and M.E., M. (2013). Clustering sentence level-text using fuzzy hierarchical algorithm. – Cité pages 29 et 43.

[Pustejovsky, 1998] Pustejovsky, J. (1998). The syntax of event structure. – Cité page 12.

[Zelizer, 1990] Zelizer, B. (1990). Achieving journalistic authority through narrative. – Cité page 39.

ANNEX

A.1 Benchmark

Event	DocID	Source	Headline
Group1	34309	FT	Barclays and ex- managers accused of crisis-era fraud
	34339	FT	Varley quits Rio role after fraud charge
	34341	FT	Advisory service agreement with Qatar plays central role
	34342	FT	Cadmium case turns toxic for Barclays
	34343	FT	The former executives caught up in investigation
	34427	WSJ	Barclays deals with a criminal dilemma
	34436	WSJ	Barclays and ex-CEO face fraud case
Group2	34410	WSJ	White House makes June ACA payments
	34408	WSJ	Senate is planning health-care vote
	34468	WSJ	Democrats dig in against health bill
	34471	WSJ	Opioid treatment would take hit, senators say
Group3	34449	WSJ Europe	Australia suspends strike operations
	34452	WSJ	U.S. downs pro-regime drone in Syria
	34425	WSJ	Skirmishing over Syria
	34435	WSJ	Trump tarnishes America's standing down under
Group4	34324	FT	US student death ensures N Korea tops talks
	34388	WSJ	U.S. pressed to act after Detainee's death
	34463	WSJ	Coroner to probe student's death
	34465	WSJ	Tour group stops taking Americans to North Korea
Group5	34315	FT	UK parliament faces 2-year Brexit battle

	34330	WSJ	Britain on its own will countfor little on the world stage
	34345	WSJ	US banks warn offund fragmentation ifhard Brexit throws up high barriers
Group6	34319	FT	Ryan risks rift with Trump over call for permanent tax reforms
	34472	WSJ	Ryan seeks momentum for tax overhaul
Group7	34336	FT	Ford shifts Focus small car production to China
	34376	WSJ	Ford to import focus from China to U.S.
Group8	34448	WSJ	What are the chances of a U.S.-China war ?
	34389	WSJ	White House seeks an ally in Beijing
Group9	34347	FT	Apple renews legal assault on Qualcomm
	34398	WSJ	Apple pushes Qualcomm lawsuit
Group10	34403	WSJ Asia	Ukraine leader meets with Trump, Pence
	34459	WSJ Europe	Ukraine leader touts U.S. ties
Group11	34390	WSJ Asia	Boeing offers bullish look ahead
	34392	WSJ Europe	Headwinds slow printer-made jet parts

TABLE A.1 – Event groups of news flow on June 21st, 2017

Event	DocID	Source	Headline
Group1	34038	WSJ	First day of Brexit talks bares divisions
	33914	FT	Hard talk London concedes to ELT's timetable as Brexit discussion begin
	33923	FT	Early Brexit talks to focus on divorce bill
Group2	33931	FT	Macron's meteoric rise lifts Europe's confidence
	34012	WSJ	Retail rebound, French vote lift European stocks
	34072	WSJ	France's Macron moment
Group3	33915	FT	Russian threat to target US forces as Syria jet shot down
	34025	WSJ	Russia cautions U.S. Over Syria
Group4	33959	FT	All eyes are on central bank shifts and whether MSCI accepts China A-shares
	34077	WSJ	Calm in markets is global
Group5	34041	WSJ	EQT deal for Rice valued at $6.7 billion
	33939	FT	EQT and Rice in $6.7bn gas producer tie-up
Group6	34022	WSJ	Cascade of violence strikes U.K.
	33977	WSJ	U.K. attack sparks outcry

TABLE A.2 – Event groups of news flow on June 20th, 2017

Event	DocID	Source	Headline
Group1	51768	FT	Brussels urges dialogue to end Catalan crisis
	51794	FT	Catalan president urges Brussels to mediate in independence clash
	51692	WSJ	Catalonia : A headache for Spain
	51744	WSJ	Catalan Leader Awaits Move by Madrid
Group2	51787	FT	Las Vegas reels from worst US mass shooting
	51788	FT	Las Vegas suffers worst mass shooting in US history
	51748	WSJ	Death toll rises to 58 in Las Vegas
	51726	WSJ	President calls Las Vegas shooting 'Act of pure evil'
Group3	51769	FT	Uber's UK head resigns as ride-hailing app battles to retain London licence
	51548	The Guardian	Uber's UK boss quits amid row in London
	51695	WSJ	Uber Executive In Europe Quits Firm
Group4	51539	FT	Help to buy is now set in cement
	51754	FT	Monarch brought down by a brutal market
Group5	51785	FT	Scientists who unwound the workings of human body clock win Nobel Prize
	51722	WSJ	Three Awarded Nobel Prize in Medicine
Group6	51700	WSJ	Google Offers a Hand to News Publishers
	51761	FT	Google offers to help news publishers sell subscriptions

TABLE A.3 – Event groups of news flow on October 3rd, 2017