FEUP at SemEval-2018 Task 5: An Experimental Study of a Question Answering System

Carla Abreu
Faculdade Engenharia
Universidade do Porto
ei08165@fe.up.pt

Eugénio Oliveira
Faculdade Engenharia
Universidade do Porto
LIACC
eco@fe.up.pt

Abstract

We present the approach developed at the Faculty of Engineering of the University of Porto to participate in SemEval-2018 Task 5: *Counting Events and Participants within Highly Ambiguous Data covering a very long tail*.[1] The work described here presents the experimental system developed to extract entities from news articles for the sake of Question Answering. We propose a supervised learning approach to enable the recognition of two different types of entities: *Locations* and *Participants*. We also discuss the use of distance-based algorithms (using Levenshtein distance and Q-grams) for the detection of documents' closeness based on the entities extracted. For the experiments, we also used a multi-agent system that improved the performance.

1 Introdution

Thousands of news articles are published every day on several media outlets. Representing and reasoning over all events in these articles is a challenging task. For instance, if we would like to answer questions about these articles like: *How many people died on the shootings in Philippi in 30th September, 2017?* or *How many people died last year on Birmingham?* or *How many people were killed by John List?*, a deep understanding is needed of many phenomena in the articles. For example, news story updates and duplicate news need to be considered in the answer processing. We can simplify the problem by identifying relevant elements from the news entities and create a structured representation to store these data.

Named Entity Recognition (NER) is a task that aims at identifying and classifying entity mentions in free text. Message Understanding Conference (MUC) defines the entities as belonging to three categories:[2] 1. *Enamex*: names, such as Locations, Persons, Organizations, and others 2. *Timex*: temporal expressions 3. *Numex*: numerical elements, such as numbers and percentages.

In this paper, we present an experimental study to extract entities from news articles to answer questions. We make use of a supervised learning approach to deal with the recognition of two different kind of entities: Locations (e.g. *Philippi, Birmingham*) and Participants (e.g. *John List*). We also have studied the use of distance algorithms (Levenshtein and Q-grams) for the near document detection based on entities extracted.

The remainder of the paper is organized as follows. In Section 2, we describe SemEval-2018 Task 5, followed by an overview of the state of the art in Named Entity Recognition in Section 3. In Section 4, we present the state of the art in the Near Document Detection task, followed by the description of the system architecture in Section 5. In Section 6, we presents the approach, followed by the experimental setup in Section 7. The results are discussed in Section 8.

2 Task Description

The main goal of SemEval-2018 Task 5 (Postma et al., 2018) is to answer questions based on a set of provided news articles, e.g. *How many killing incidents happened in 2016 in Columbus, Mississippi?*. Each question has three components: an event type and two event properties. Each question contains one out of four event types: *killing*, *injuring*, *fire burning*, and *job firing*. Event Properties are all the related characteristics associated with the event. They can include *Locations* (City or State), *Participants* (First Name, Last Name, Full Name), and *Time* (Day (e.g. 1/1/2015), Month (e.g. 1/2015) or Year (e.g. 2015)). There are three

[1] https://competitions.codalab.org/competitions/17285

[2] http://afner.sourceforge.net/what.html

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 667–673
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

subtasks:

- Subtask 1 (S1): Find the single event that answers the question
- Subtask 2 (S2): Find all events (if any) that answer the question
- Subtask 3 (S3): Find all participant-role relations that answer the question

3 Named Entity Recognition

A wide range of approaches have been developed to tackle NER. Early systems deal with this issue by making use of handcrafted rule-based algorithms (Hearst, 1992). More recently, systems focus on machine learning techniques (supervised learning (Florian et al., 2003), semi-supervised (Collins and Singer, 1999; Mikheev et al., 1999), and unsupervised learning). However, the major drawback of supervised learning is its dependence on annotated data. In the case of unavailability of training examples, handcrafted rules remain the practical technique (Riaz, 2010).

4 Near Document Detection

In the large amount of news articles that are published every day, the same information can be repeated in many different articles. The identification of similar or near-duplicate documents is applied in: plagiarized documents detection (Hoad and Zobel, 2003), similar web pages detection (Henzinger, 2006), and similar news articles detection (Abreu et al., 2015).

Identification of similar or near-duplicate pairs of documents in a large collection is a significant problem with wide-spread applications. Kumar and Govindarajulu (2009) present approaches used to solve this issue. For those kind of problems, three main approaches are proposed: based on URLs, on lexicon and, the third and more sophisticated, on semantics (Abreu et al., 2015).

In the work presented here, we are using the semantics-based approach applied to the information previously extracted from the news articles.

5 Architecture

The system consists of the following main components:

Creating a Structured News Representation.

Copyright 2017 by WJXT News4Jax - All rights reserved.
To get alerts for breaking news, grab the free NBC4 News App for iPhone or Android.
contact kimber laux at klaux@reviewjournal.com or 702 - 383 - 0283.
Contact Jessica Terrones at jterrones@reviewjournal.com or at 702-383-0381.

Table 1: Journalistic Patterns

D	M	Y	Regular Expression
x	x	x	(Jan. [1-9]+[1-9]*, [1-2][0-9][0-9][0-9])
	x	x	(December [1-2][0-9][0-9][0-9])
		x	[1-2][0-9][0-9][0-9]

Table 2: Temporal Regular Expressions

Figure 1 presents the architecture used to parse the news article. After converting CoNLL to plain text, journalists patterns are removed as demonstrated in Table 5. Journalistic patterns could be relevant for the reader, but not for the entity recognition task. The output of this system is a structured news representation with a list of Event Types, Locations, Participants, and Temporal Expressions. Additionally, the following sources of information are also extracted: the news identifier, publication date, and news title. To result in this representation, the following four extractions are performed:

Extract list of Event Types. We use WordNet (Fellbaum, 1998) to create a list of words that can be used to describe an event type. Our approach uses the news article title and body for the event type recognition. For each one of these elements, the English Snowball Stemmer is applied. We consider a document to have a certain event type if at least one term that describes an event type is present in the news title or body.

Extract Locations and Participants. For the Locations and Participants recognition, a supervised approach is used. The approach proposed is described in Section 6.

Extract Temporal Expressions. Our approach to finding temporal information in the news article is based on the application of regular expressions. Table 2 presents some of the regular expressions used.

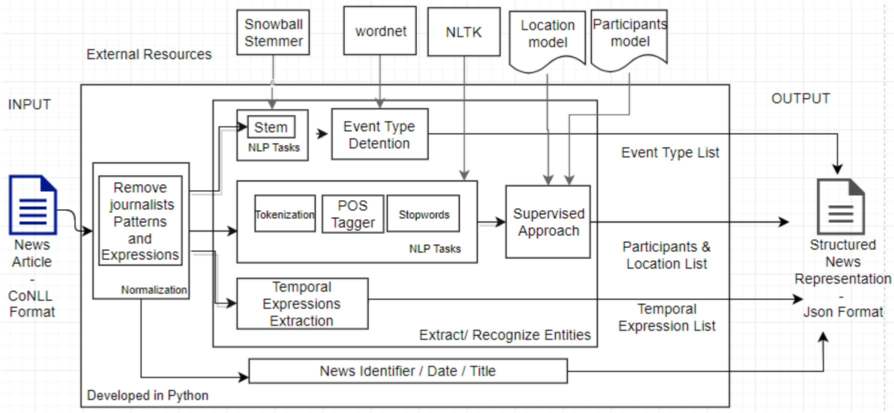


Figure 1: Create a structured news representation approach

Extract Auxiliar Information. The title, publication date, and news identifier are also extracted from the news article to create the structured news representation.

Search all the news that answer a question

When the system receives a question, an answer will be retrieved based on the structured news representation. Firstly, for each element (Event Type and Event Propreties) a list of news articles that has some relation with the element under analysis is composed. In the end, the news or set of news articles that address all the items under analysis are extracted.

Near document detection

The near document detection was done based on the set of news that answers a question. The approach is explained in Section 6.

Counting participants

Similarly to what happened in the case of previously mentioned events' extraction, this one only uses a news article or a set of news articles that answer a question. For this set of news articles, we only process the information given by the news article title. For each Event Type we manually define the variation trend (increase/ decrease/ stable) - e.g. the number of death can increase with the decrease of the number of injured - in a killing event. We started this process by normalizing and removing temporal expressions from the news article title. After, we applied the POS-tagger and split the sentence into subsentences separated by comas. We started to recognize the event type for each subsentence. When we found it, we checked if the subsentence also includes a numerical element ('CD' - Post tagger) - this element is considered as a number of participants associated with the event type. Once extracted the number of participants associated with each news article, we connect this information with news article's date. Finally, we try to find the maximum or the minimum of participants depending on the temporal event type trend.

The system we are presenting here was developed in Python 2.7. It uses some python libraries: Natural Language Toolkit (NLTK) - Wordnet, English Snowball Stemmer, Stopwords, POSTagger; Python multi-Agent Development Environment (SPADE); Scikit-Learn - tree, RandomForestClassifier, ExtraTreesClassifier, LinearSVC; Json; and, Regular Expressions (re).

6 The proposed approach

6.1 NER Supervised Approach

In this subsection, we describe the implementation details of the proposed approach for recognizing Locations and Participants.

Natural Language Processing Tasks:

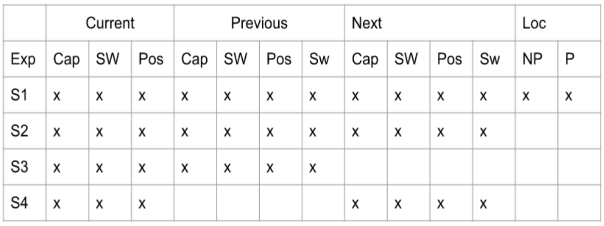
The data was preprocessed with two NLP tasks: part of speech tagging and stop word recognition.

Features

Supervised learning techniques require their input to be categorized. When extracting information from news documents, it is common to label each word with a set of features. These features allow the SL approach to recognize an entity in a given document. We extracted the following features: 1. CAP (Capitalized) indicates whether a word: contains no capital characters,

	shooting	at	a	west	Phoenix	apartment	that	left	one	man	dead
CAP	0	0	0	0	1	0	0	0	0	0	0
PT	D	NN	IN	DT	NNP	NN	WDT	VBD	CD	NN	NN
SWI	0	1	1	0	0	0	1	0	0	0	0
SWA		at	a				that				

Table 3: Categorizing each word on a sentence



	Current			Previous				Next				Loc	
Exp	Cap	SW	Pos	Cap	SW	Pos	Sw	Cap	SW	Pos	Sw	NP	P
S1	x	x	x	x	x	x	x	x	x	x	x	x	x
S2	x	x	x	x	x	x	x	x	x	x	x		
S3	x	x	x	x	x	x	x						
S4	x	x	x					x	x	x	x		

Figure 2: Features used in each scenario

has only its first letter capitalized, or all its characters are capitalized; 2. PT (POS Tagger Association)[3] identifies the part of speech tag of a word, such as noun, verb, adjective, etc.; 3. SWI (Stop–Words Identification) indicates whether a word is a stop-word; 4. SWA (Stop-Words Association) associates a corresponding stop-word; 5. NP - Paragraph records a numeric identifier of the paragraph in which the word appears.

Table 3 presents example of features computed for each word in the phrase “shooting at a west Phoenix apartment that left one man dead”. For instance, the word “Phoenix” is capitalized ($CAP = 1$), corresponds to a noun ($PT = NNP$), and is not a stop-word ($SWI = 0$). Note that we aggregate all sequential capitalized words as one, e.g. “Salt Lake City” will be combined in a single word to be classified.

We believe that a simple association as illustrated in Table 3 is not enough to categorize a word for the Named Entity Recognition task. For this reason, we also consider the word context in the document, i.e., the current word (C), the previous word (P), and the next word (N). Here we indicate the word position following the feature abbreviation, e.g., “C CAP” indicates whether the current word is capitalized or not.

Data Cleaning and Transformation

Data quality is the main challenge of information management. To guarantee data quality, two processes were executed: data cleaning and data transformation. Tables 4 and 5 present the data transformation for POS tags and stop-words. Stop-words have no value for SWA. To fix this, we replace an empty value by the character “X” and we encode this value as demonstrated in Table 5.

PostTagger	Rep
DT	0
NN	1
NNP	2
VBD	3
...	...

Table 4: POSTtagger

Stop-Word	Rep
X	0
a	1
that	2
and	3
...	...

Table 5: Stop Words

Classification Algorithms

Supervised learning techniques create a model that predicts the value of a target variable based on a set of input variables. One challenge is to select the most appropriate algorithm for the task of classifying Locations and Participants. We have compared the following algorithms: Support Vector Classifier (SVC); Decision Tree Classifier (Tree); Random Forest Classifier (Random); Extra Trees Classifier (Extra). As demonstrated on Table 6, different configurations were attempted for each algorithm. Implementations of these algorithms are provided by the Python library scikit-learn library[4].

6.2 Near Document Detection

The answer to a question in this SemEval task consists of the following: question identifier, set of the news articles that help to answer the question, and a numerical answer.

The numerical answer of a question is dependent on the question task. Task 2 requires a number of unique events that correspond to a question. For this purpose, it is essential to detect similar news documents within the given set. To detect similar documents, we use the structured news representation described above. Each pair of news articles is compared based on: their titles, their lists of Participants, and their lists of Locations.

7 Experimental Setup

7.1 NER Approach

Data Resources

The SemEval 5 competition provides data for the purpose at stake. The data made available in this competition is a set of English news articles. To extract locations and participants from crime

[3] (POSTagger - All Tags) - http://www.nltk.org/book/ch05.html visited on 2017, November

[4] http://scikit-learn.org/stable/, visited in November 2017

Alg/ID	Configuration
SVC 1	Default scikit learn configuration
SVC 2	kernel=”linear”
SVC 3	kernel=”sigmoid”
Tree 1	Default scikit-learn configuration
Tree 2	criterion=”gini”, splitter=”best”, min samples split=2
Tree 3	criterion=”entropy”, splitter=”best”, min samples split=2
Tree 4	criterion=”entropy”, splitter=”random”, min samples split=2
Tree 5	criterion=”gini”, splitter=”random”, min samples split=2
Tree 6	criterion=”gini”, splitter=”best”, min samples split=4
Tree 7	criterion=”entropy”, splitter=”best”, min samples split=4
Random 1	criterion=”gini”, n estimators=10
Random 2	criterion=”gini” n estimators=5
Random 3	criterion=”gini”,n estimators=20
Random 4	criterion=”entropy”, n estimators=10
Random 5	criterion=”entropy”,n estimators=5
Random 6	criterion=”entropy”,n estimators=20
Extra 1	criterion=”gini”, max features=”auto”
Extra 2	criterion=”entropy”, max features=”auto”
Extra 3	criterion=”gini”, max features=”sqrt”
Extra 4	criterion=”entropy”,max features=”sqrt”
Extra 5	criterion=”gini”, max features=”log2”
Extra 6	criterion=”entropy” max features=”log2”
Extra 7	criterion=”gini” max features=None
Extra 8	criterion=”entropy”, max features=None

Table 6: Classification Algorithm Configurations

news, additional annotations were done. A set of 10,580 individual words were annotated in three categories: Locations, Participants, and Others - where all the sequential capitalized words were aggregated as one, e.g., in “as she left Jackson Memorial Hospital”, the annotated elements are: [as], [she], [left],[Jackson Memorial Hospital].

Evaluation

The evaluation metrics used to evaluate this approach are Precision (P), Recall (R), and F1 (F). Due to a large number of experiences and in order to correctly analyze the obtained results, we made use of a multi-agent architecture to find the best results. For this evaluation, we defined a utility function and we introduced an auction mechanism to enable some kind of negotiation. This mechanism is based on English auction, where each agent can propose their bids following the auction requirements. Our agents represent the different configuration of the classification algorithms and each bid reveals their result on a specific test scenario. We expect that in this experiment recall is the most important metric, thus it is assigned a higher weight than the other metrics. Our utility function was defined as follows:

$$U = 0.5 * R + 0.25 * P + 0.25F1$$

In order to reduce the data to be analyzed we exclude all combinations with low performance, namely all combinations where either Recall, Precision, or F1 has a mean value bellow 60% or a standard deviation above 15%.

Experiments

A supervised learning system was needed to generate a model. The classification algorithms and the scenarios (S1, S2, S3, and S4) defining values of features are those described in section 6.1.

Our experiments were done taking cross-validation with $k = 7$ into account. We divided the annotated data into partitions of training data (75%) and testing data (15%).

7.2 Near Document Detection

Data Resources

Near document detection approach was studied with the dataset provided at the end of the competition. Each intended answer includes a list of similar documents identified in the given dataset and aggregated according to the corresponding question. For each answer, we created a script to aggregate all news articles in pairs. Additionally, a label indicating whether a pair is similar or nor (pairs that are contained in the same set are similar) was added. In total, this resulted in 61,931 pairs of news articles.

Evaluation

We evaluate the performance of various thresholds on near document detection by applying the metrics: Precision, Recall, and Accuracy.

Experiments:

For each pair of news articles, we have calculated the similarity between their elements: title (T), list of participants (Part), and list of locations (Loc). For the sake of comparison, we have used two distance algorithms: Levenshtein (L) (Levenshtein, 1966) and Qgrams (Q) (Ullmann, 1977). We defined two scenarios (SS1, SS2), differing in the weights of the document elements as follows:

$$SS1 = 0.50T + 0.25Loc + 0.25Part$$

$$SS2 = 0.34T + 0.33Loc + 0.33Part$$

Exp	Alg	P	R	F	U
S3	Tree 6	66.47	70.94	68.34	69.17
S1	Extra 8	71.67	67.63	69.13	69.02
S1	Tree 2	71.28	67.53	69.08	68.85
S1	Tree 3	70.53	67.79	68.90	68.75
S2	Tree 4	70.35	67.55	68.54	68.50

Table 7: Recognizing Participants - Results

Exp	Alg	P	R	F	U
S3	Tree 3	70.32	66.68	68.13	67.95
S1	Extra 8	68.91	67.53	67.97	67.98
S1	Extra 4	68.13	67.40	67.64	67.64
S1	Extra 7	67.11	70.06	68.34	67.16
S1	Extra 2	68.28	64.84	66.41	66.09

Table 8: Recognizing Location - Results

8 Analysis and Results

8.1 NER Approach

Due to the large volume of combinations and their corresponding results, we used a multi-agent system to simplify the analysis. Tables 7 and 8 present the best 5 results achieved on extracting Participants and Locations respectively. We considered the results only from two algorithms: Decision Tree and Extra Tree Classifier. Both approaches show that context helps the recognition task.

8.2 Near Document detection

Table 9 presents the results achieved for various threshold values. Changing the threshold value causes small variations in the performance of the Qgrams algorithm, but large variation in the performance of the Levenshtein distance algorithm. The scenarios presented here are not sufficient to determine if two news articles are similar or not. These results indicate that in cases where news articles refer to the same subject, a reduced news article representations is not sufficient to distinguish different events.

Alg	Function	Threshold	P	R	A
L	SS1	75	12.24	12.86	86.56
L	SS1	80	6.97	60.86	36.22
L	SS1	85	8.26	39.98	62.22
L	SS2	75	13.46	11.96	87.64
L	SS2	80	7.97	40.79	60.30
L	SS2	85	9.37	16.05	82.21
Q	SS1	75	12.30	12.86	86.60
Q	SS1	80	13.38	10.98	88.00
Q	SS1	85	13.38	10.97	88.00
Q	SS2	75	13.61	10.70	88.21
Q	SS2	80	13.23	11.07	87.89
Q	SS2	85	13.38	10.98	88.00

Table 9: Near Document Detection Results by Threshold

8.3 SemEval Results

SemEval-2018 Task 5 contains 3 subtasks, on which we achieved F1 score of 24.65, 30.51, 26.79 respectively.

9 Conclusion and Future Work

In this work, we presented an experimental study that addresses the Question Answering challenge in SemEval-2018 task 5. We have used Named Entity Recognition approaches to identify entities such as Location, Participants, Temporal Expressions, and Event Types. We used a structured news representation to perform the required tasks: 1. to answer questions on counting events 2. to detect which distinct documents provide an answer to a question; and 3. to answer questions by counting event participants.

The use of multi-agent system was crucial in order to find the best performing algorithm. Our utility function allowed us to have a previous definition of the influence of each evaluation metric on the overall evaluation. The resulting system can be applied to other scenarios by adapting the utility function according to their requirements. In the future, our system can be improved to include multiple combinations (e.g., on the near document detection we can use a different combination of elements and algorithms).

Task 2 has been solved with extracting information. Future work for this task could include another study of a supervised learning approach that is based on the entire information available in a news article. However, such a requires a corresponding annotated corpus. Our approach to Task 3 was relatively naive, since it does not consider the relationships between entities (participants and event type). Future work should investigate a more elaborate approach.

Event type behavior should also be studied, as we believe that some events could present temporal trends. For instance, we expect to observe an increase of the number of deaths/injuries described in crime news documents over time.

References

Carla Abreu, Jorge Teixeira, and Eugénio Oliveira. 2015. Encadear: Encadeamento automático de notícias. *Oslo Studies in Language*, 7(1).

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In

1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Monika Henzinger. 2006. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 284–291. ACM.

Timothy C Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarized documents. *Journal of the Association for Information Science and Technology*, 54(3):203–215.

J Prasanna Kumar and P Govindarajulu. 2009. Duplicate and near duplicate documents detection: A review. *European Journal of Scientific Research*, 32(4):514–527.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics.

Marten Postma, Filip Ilievski, and Piek Vossen. 2018. Semeval-2018 task 5: Counting events and participants in the long tail. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. Association for Computational Linguistics.

Kashif Riaz. 2010. Rule-based named entity recognition in urdu. In *Proceedings of the 2010 named entities workshop*, pages 126–135. Association for Computational Linguistics.

Julian R. Ullmann. 1977. A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. *The Computer Journal*, 20(2):141–147.