Vol.4,No.1, July, 2015
168 - 177,ISSN:0975-5446 [VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY]

Hybrid Approach for temporal information processings

PATEL Parul
M.Sc(I.T) Programme
Veer Narmad South Gujarat University
parul.patelns@gmail.com

PATEL S.V.
Department of Computer Science
Veer Narmad South Gujarat University
patelsv@gmail.com

Abstract

Using temporal information available in the document and exploring search results based on timeline is an important requirement of today's search engines. To do so, the first step is to extract and process the time related information available in the document and make it explicitly available for use by search engine. Processing such temporal expression demands accurate temporal tagger which not only processes temporal expression but also translates it into standard form which can be used for further processing. Many temporal taggers have been developed by researchers, however they have limited capabilities to extract temporal expressions like festivals. Further, majority of taggers are developed for extracting expressions from western documents where such festival occur on fixed date. In Indian documents, most of the festivals have dynamic dates year on year. We have developed a versatile tagger which overcomes such limitations. Further, instead of using only rule based approach, we have used machine learning approach for temporal expression recognition and rule based approach for normalization.

Keywords: temporal expression, temporal expression recognition, CRF, machine learning,

1. Introduction

Time is very important dimension in any information retrieval system. Temporal information is present into the form of temporal expression in any document. Processing such temporal expression from raw text is fundamental requirement for application like text summarization, question answering. Our long term goal is to use this temporal information in exploring search results on timeline. A temporal expression also known as Timex also refers to every natural language phrase that denotes a temporal entity like interval or an instant. For e.g. " Prime Minister Narendra Modi will visit china tomorrow.", "India won the test match on last Friday". Phrase "tomorrow", "last Friday" denotes timexes. Such temporal expressions can be classified into following categories[1]:

Fully Qualified: A temporal expression is fully qualified with respect to the binding when all the information required to infer a point in the time domain are fully included inside the expression. In this category the following expressions falls: March 15 2010, 31st January 1984 or 14/11/2010. It is easy to detect fully qualified temporal expressions because of their rigid lexical form.

Deictic: In these type of expressions, it is required to take into account reference time (when document has been written or when dialog has been made or speech has been given). Such temporal expressions could not be properly allied to the time without reference information . For e.g. five months ago, tomorrow, today, two week before, before last Christmas.

Anaphoric: These expressions can be mapped to a precise point in the time domain only taking into account temporal expressions previously mentioned in the text or during the speech. Examples of this category are: March 15, the next week, Saturday. The only difference between deictic and anaphoric expressions is the location of the temporal reference: for deictic expressions it is the time of utterance or publication, for anaphoric expressions it is a time previously revoked in the text or speech. Anaphoric expressions constitute a future challenge for the scientific research in this field.

Vague: Some temporal expressions represents vague temporal information and it is difficult to precisely normalize into some absolute value. For e.g. few days ago, few months later, later, in several weeks etc.

Usually there are two common methods for extraction and processing of above temporal expressions. (1) Machine learning Approach (2) Rule based approach. Rule based approach gives very good accuracy, but developing a set of rules is very time consuming and requires more human efforts . Moreover it is difficult to add new rules and maintain consistency with previous rules. In contrast, machine learning is very effective method, but require good feature set and training data for learning process. We have combined both approaches for temporal expression recognition and normalization to take advantage of both methods. We have used CRF (Conditional Random Fields) based machine learning approach for recognition and rule based approach for normalization.

The Rest of the paper is organized as follows: Section 2 describes literature review of various existing temporal taggers. In Section 3 system architecture of our tagger is presented. In Section 4, evaluation of our temporal tagger with different corpus and analysis is described. Section 5 includes conclusion and future work.

2. Literature Review:

The Message Understanding Conferences (MUCs) in 1996 and 1998 have played a significant role, but their evaluations covered only recognition of TEs, while a novel contribution towards the normalization of TEs was made in 2000[2]. GUTime was a rule based system which was developed an extension of TempEx tagger. It was based on TimeML TIMEX3 format, which allows a functional style of encoding offsets in time expressions. It was evaluated on TERN 2004 corpus and achieved 85% of F-measure. Llorens has developed temporal information extraction system based on CRF for Spanish documents with F-measure of 91%[3]. KUL is a machine learning based system for recognition and normalization of temporal expression with 0.85% precision and recall of 0.84%[4]. Negri and Merseglia has developed a rule based system which involves tokenization, part-of-speech tagging based on a list of 5000 entries retrieved from WordNet. Then, the recognized text is processed by a set of approximately 1000 basic rules. Recognized temporal expressions and information around that is used for normalization. Then

composition rules are used to resolve ambiguities wherever multiple tag placements are possible. The results in terms of F-measure on ACE 2004 data are 92.6%, 83.9%, 87.2% for detection, recognition and determining the VAL attribute value, respectively [5]. Heideltime is high quality rule based tagger for temporal expression recognition and normalization with 0.90% precision and 0.82% recall [6]. The Yamcha is machine learning based tagger which uses SVM and FOIL for chunking and classification of chunks. They got precision of 80.05% , recall of 73.71% and F-measure of 76.75%. They have concluded that use of SVM leads to overfitting[7]. Jelena has developed a system for temporal information extraction and interpretation for serebian language with precision of 0.93%, recall of 0.96% and F-score or 0.94%[8]. SUtime is the library for recognizing and normalizing temporal expressions developed by stanford university. It is rule based system developed in java [9]. SuTime is having limited support for range queries.

It has been observed that many of the above tools do not support festivals as temporal expressions and some of them support only fixed day festival like Christmas. We have analyzed Indian news documents, where we have found Indian festivals which does not have fix date.(e.g. Diwali, Ramzan etc.) every year. Above tools do not support such variable date festivals and most of them are not easily extensible. We have developed a hybrid approach for temporal expression recognition and normalization which eliminates above limitations as well as it is generic enough for further extension in future.

3. System Architecture

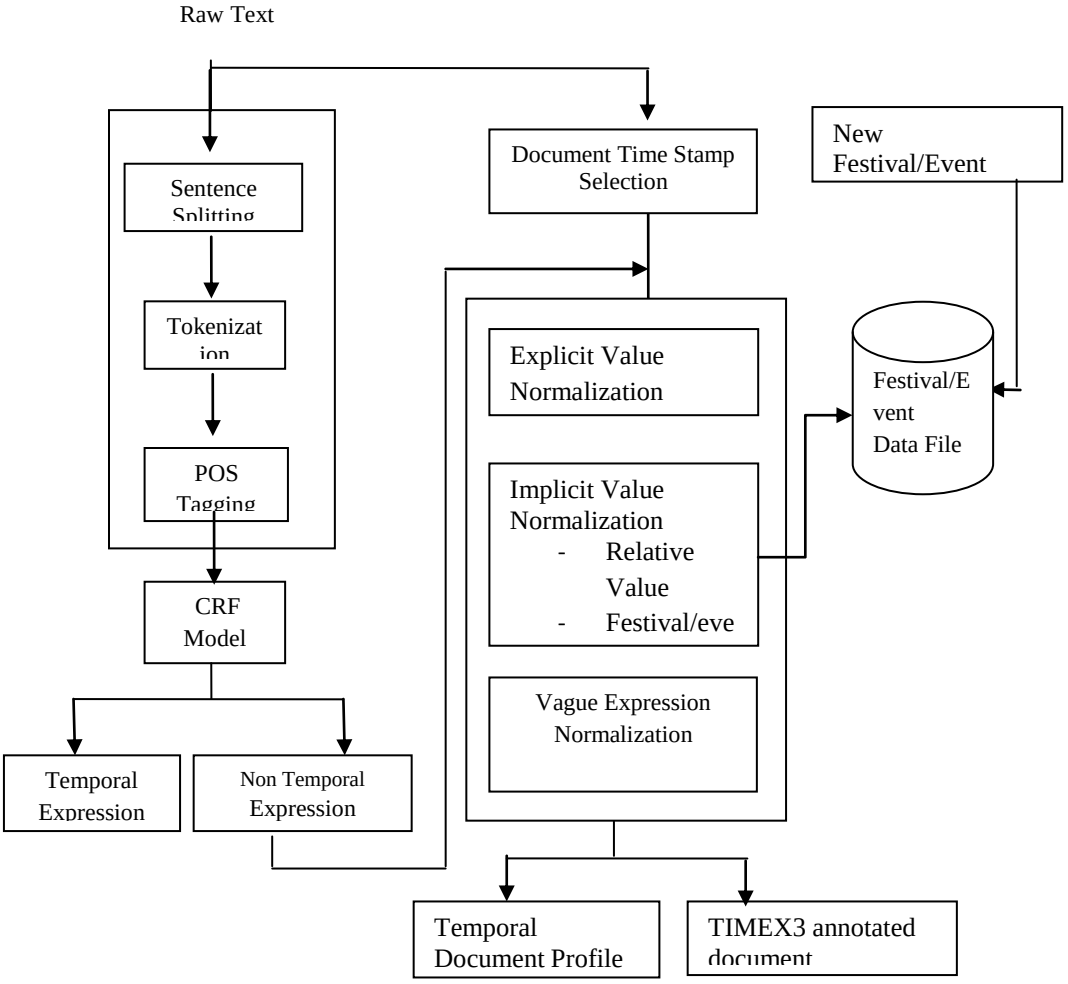

Figure 1: Block diagram of Temporal Information Processor

3.1 Recognition of temporal expressions

We have used CRF based machine learning approach for recognition of temporal expressions. For recognition, we have used following steps:

3.1.1 Training and Testing Data Preparation for Recognition

We used WikiWar data set which contains 22 Xml historical documents containing different type of temporal expressions. WikiWar data set contains total of almost 1,20,000 tokens and 2671 temporal expressions annotated in TIMEX2 format [9]. We used 17 documents from the whole collection as the training data set and 5 documents as the testing data set. We have selected 163 timex2 annotated documents from Tempeval corpus. As Temporal expressions present into the wikiwar and Tempeval do not focus on any festivals, we used 30 other manually annotated news documents which contain Indian festivals and other temporal expressions combinations. We have used such 20 manually annotated news documents for testing of recognition module. Overall , we have used 185 documents for training. Data cleaning was performed on all training/testing documents to remove Xml tags except the temporal tags. Then all documents were preprocessed to convert data into a form for further processing of training/testing. Following steps are performed on documents.

- **Sentence Splitting**

Stanford sentence splitter was then used to split all sentences of all documents.

- **Tokenization**

Each sentence is then split into tokens with their respective position in the sentence by using Stanford tokenizer.

- **Pos Tagging**

Stanford POS tagger was then applied to extract all part of speech (POS) features of each token .

3.1.2 Selection of Suitable machine learning algorithm

A machine learning technique that has recently been introduced to tackle the problem of labeling and segmenting sequence data is Conditional Random Fields. Unlike Hidden Markov Models, CRFs are based on exponential models in which probabilities are computed based on the values of a set of features induced from both the observations and label sequences. This enables the incorporation of overlapping and interacting features into the model. CRFs have been shown to perform well in a number of natural language processing applications, such as POS tagging [11], shallow parsing or noun chunking [12], and named entity recognition [13]. Their characteristics make CRFs ideally suited for the specific task of recognizing timexes as they provide us with a framework for combining evidence from different sources to maximize performance. W. Cohen used the implementation of CRFs from the minor- Third toolkit for extracting timexes from text [14]. We have selected CRF classifier for recognition of temporal expressions.

3.1.3 CRF Feature builder

We extracted Months, days and year features from the tokens. These tokens are helpful in identifying the temporal expressions. This way each token will have two features extracted namely a) Calendar features that involves months, days and year and b) POS features. We have considered following temporal expressions :

- A List of periodic temporal set: Hourly, Daily, Weekly, Monthly, Yearly
- A List of Seasons: Spring, Winter, Monsoon, Summer etc.

- A list of relative days: Yesterday, Today, Tomorrow etc
- A list of all Indian festivals occurring on fixed days: Independence Day, Republic Day, Teacher's Day, Gandhi Jayanti etc.
- A list of all Indian Festivals occurring on variable days: Diwali, Holi, Navratri, Women's Day, Rakhi, Durgashtami etc.
- A list of months: January, February...December.
- A list of temporal expression modifier: Last, This, Mid, Recent, Earlier, Beginning, Late
- A list of decades: twenties, thirties etc.
- A list of Week Days: Monday, Tuesday……...Sunday etc.

3.1.4 CRF model building and classification

The CRF feature builder has generated features for the CRF machine learner. The context window for the CRF was set to be five words. Each pair of temporal expressions at the unigram sentence level is used for training. We have used CRF++ 0.58, an open source implementation of the conditional Random Field (CRF) machine learning classifier for our experiments. We have used CRF++ templates to capture the relations between different features to recognize temporal expressions. Template file contains unigram template which defined number of tokens to be considered for prediction of token as temporal expression. The training set is prepared which is given as an input to the CRF algorithm. The CRF algorithm learns from the training samples and gives a model. The training set contains a) token, b) its calendar feature, c) POS tag and the d) label which is either temporal expression or non-temporal expression.

Table 1: Input Data

Input		
He	OTH	PRP
Had	OTH	VBD
Handled	OTH	VBN
distribution	OTH	NN
of	OTH	IN
Cheques	OTH	NNS
Worth	OTH	JJ
Crores	OTH	NNS
during	OTH	IN
The	OTH	DT
Khoraj	OTH	NNP
Land	OTH	NN
acquisition	OTH	NN
Drive	OTH	NN

In	OTH	IN
Sanand	OTH	NNP
in	OTH	IN
December	CAL	NNP
last	CAL	JJ
Year	CAL	NN
.	OTH	.

Table 2 : Output Data

Output			
He	OTH	PRP	nonTemporalExp
Had	OTH	VBD	nonTemporalExp
Handled	OTH	VBN	nonTemporalExp
Distribution	OTH	NN	nonTemporalExp
of	OTH	IN	nonTemporalExp
Cheques	OTH	NNS	nonTemporalExp
Worth	OTH	JJ	nonTemporalExp
Crores	OTH	NNS	nonTemporalExp
during	OTH	IN	nonTemporalExp
The	OTH	DT	nonTemporalExp
Khoraj	OTH	NNP	nonTemporalExp
Land	OTH	NN	nonTemporalExp
Acquisition	OTH	NN	nonTemporalExp
Drive	OTH	NN	nonTemporalExp
In	OTH	IN	nonTemporalExp
Sanand	OTH	NNP	nonTemporalExp
in	OTH	IN	nonTemporalExp
December	CAL	NNP	temporalExp
last	CAL	JJ	temporalExp
Year	CAL	NN	temporalExp
.	OTH	.	nonTemporalExp

3.2 Normalization

Once recognition process is completed, all recognized temporal expressions are passed to normalization module. We have used traditional rule based approach for normalization of temporal expression recognized by CRF recognizer. In this phase, all temporal expressions are assigned their possible absolute value. First all explicit temporal expression are normalized to their absolute values by selecting their appropriate function. We have created a database which contains dates of all Indian festival of last 50 years and next 25 years. In addition, it contains all popular days with their corresponding date. Then all relative temporal expressions are assigned their absolute value by taking document timestamp as reference. All relative temporal expressions that belong to some popular day, event or festival are normalized to their respective absolute values by extracting year from timestamp and searching respective date from database. For e.g. consider following example with timestamp 12-05-2014.

(1) Narendra Modi visited Gandhi Ashram on this republic day.
Narendra Modi visited Gandhi Ashram on < TIMEX3 tid="t1" type="DATE" value="2014-01-26" > this republic day < /TIMEX3 >.

(2) Narendra Modi Will be in Srinagar on this Diwali &"I will spend the day with our sisters & brothers affected by the unfortunate floods" PM Modi tweeted.

Narendra Modi Will be in Srinagar on <TIMEX3 tid="t1" type="DATE" value="2014-10-23" > this Diwali </TIMEX3> &"I will spend the day with our sisters & brothers affected by the unfortunate floods" PM Modi tweeted.

During recognition phase some temporal expressions are recognized which are difficult to normalize to some specific absolute value. For e.g few years back, in early days, few weeks later. Our CRF classifier recognizes all kind of temporal expressions but normalization module is not normalizing such vague expressions into absolute value.

3.3 Intermediate representation of the temporal expressions

After Normalization phase, extracted temporal expressions with their normalized value are stored into vector called temporal document profile which can be used further in application like time line generation, question answering, document summarization. Temporal document profile contains all temporal expressions present into the document and their corresponding normalized value. It contains starting position of the sentence in the document where temporal expression is present and ending position of the sentence. This positions can be used to extract sentences where temporal expressions occur in the document.

3.4 Timex3 Annotation

After Normalization phase, it generates annotated file in which all temporal expressions are annotated as per TIMEX3 Standard. It includes Timex Id, Type and Value attribute of TIMEX3. Timex Id uniquely identify temporal expressions within the document. Type attribute is assigned one of the four values: DATE, TIME, DURATION and SET. DATE describes calendar time. For. e.g 'Friday', 'October 1, 1996' ,' the second of November',' yesterday',' in October of 1953'. TIME expression defines time of the day. For e.g. '5 minutes ago', 'two minutes later', 'ten minutes to five'. DURATION expression describes explicit durations like '2 months','five years',''two days before'. SET expression defines set of times, for e.g. 'twice a week','four days a week','every day' etc. Following example gives look towards annotation file and temporal document profile.

Speed News » **13 Feb 2014 IST**
Those who indulge in hooliganism and eve teasing **on Valentine day**, beware. Ludhiana police claims to go very hard against such elements and have put in fool proof arrangement to deal with such people.Doctors from the oncology department of Government Medical College and Hospital (GMCH) and volunteers of CanKids will be celebrating Valentine's Day by organizing a blood donation camp. It will be held at the promenade of Futala lake. Nearly 150 supporters in saffron were on roads in a 'Chetna rally' to convey the message. In a prelude to discourage western culture in state, supporters of Bajrang Dal- youth wing of the Vishwa Hindu Parishad- have on Wednesday warned people against celebrating Valentine's day or indulging in any indecent act on February 14. Despite global economic slowdown, companies are ready to cash in on Valentine Day's and an estimated Rs 18,000 crore with record sales of popular gift articles like flowers, jewellery, chocolates, toys, readymade garments, mobile phones, electronic gadgets, wrist watches, according to ASSOCHAM based on feedback collected from producers or manufacturers.

Figure 2: Example of data from Times of India

Speed News »
<TIMEX3 tid="t1" type="TIME" value="1985-02-13">13 Feb 2014,</TIMEX3> IST
Those who indulge in hooliganism and eve teasing on <TIMEX3 tid="t2" type="DATE" value="1986-02-14">Valentine day</TIMEX3>, beware. Ludhiana police claims to go very hard against such elements and have put in fool proof arrangement to deal with such people.Doctors from the oncology department of Government Medical College and Hospital (GMCH) and volunteers of NGO CanKids will be celebrating <TIMEX3 tid="t3" type="DATE" value="1986-02-14">Valentine's Day</TIMEX3> by organizing a blood donation camp. It will be held at the promenade of Futala lake.Nearly 150 supporters in saffron were on roads in a 'Chetna rally' to convey the message. In a prelude to discourage western culture in state, supporters of Bajrang Dal- youth wing of the Vishwa Hindu Parishad- have on <TIMEX3 tid="t7" type="DATE" value="1985-09-18">Wednesday</TIMEX3> warned people against celebrating <TIMEX3 tid="t8" type="DATE" value="1986-02-14">Valentine's day</TIMEX3> or indulging in any indecent act on <TIMEX3 tid="t9" type="DATE" value="1986-02-14">February 14</TIMEX3>. Despite global economic slowdown, companies are ready to cash in on <TIMEX3 tid="t11" type="DATE" value="1986-02-14">Valentine Day </TIMEX3>'s and an estimated Rs 18,000 crore with record sales of popular gift articles like flowers, jewellery, chocolates, toys, readymade garments, mobile phones, electronic gadgets, wrist watches, according to ASSOCHAM based on feedback collected from major producers or manufacturers.

Figure 3: Annotated File as per TIMEX3 Attribute

Table 3: Temporal Document Profile

Temporal Expression	Normalized Value	Start Position	Date	Month	Year
13 Feb 2014	2014-02-13	15	13	2	2014
Valentine day	2014-02-14	86	14	2	2014
Valentine's Day	2014-02-14	280	14	2	2014
Wednesday	2014-02-12	450	12	2	2014
Valentine's Day	2014-02-14	480	14	2	2014
February 14	2014-02-14	525	14	2	2014
Valentine Day	2014-02-14	596	14	2	2014

4. Evaluation and Analysis:

For evaluation, we have selected 3 datasets. Each of them are rich of temporal expressions. DataSet1 is Wikiwar corpus which contains approximately 2671 temporal Expressions. DataSet2 is collection of biographies of well known person of india. Dataset 3 is Tempeval corpus. Dataset

4 is a collection of 50 news articles of different time period. This data is collected from Times of India. We have got following accuracy in recognition and normalization module.

Table 4: Results of recognition on various corpus

	Precision	Recall	F-Measure
DataSet1	92.84%	99.63%	96.11%
DataSet2	90.72%	94.34%	92.49%
DataSet3	98.90%	97.24%	98.06%
Dataset4	91.40%	94.54%	92.94%

Table 5: Results of normalization on various corpus

	Precision	Recall	F-Measure
DataSet1	90.03%	91.35%	90.68%
DataSet2	91.40%	94.43%	91.87%
DataSet3	88.90 %	96.54%	92.56%
Dataset4	91.32%	95.67%	93.44%

Resolving relative temporal expressions is more challenging. We have considered document creation date as reference date for normalization of any relative temporal expressions. But in some cases, it is not appropriate to user DCT as reference date. For e.g. 'He concluded the 2012 annual general meeting by saying " The next year will be very important year for them' . In this example, next year should be normalized using previous sentence da te instead of DCT. Our recognizer module has recognized all different types of temporal expressions, but it is difficult to normalize all of them to absolute values. For e.g. vague temporal expressions like 'in several day', 'few months later','in the evening' are difficult to normalize. Our system is able to recognize it but normalization is not handled . Durative expressions are recognized separately for e.g from October 2000 to 2005 are considered as October 2000 and 2005.

5. Conclusion & Future Work:

The paper presented a system with hybrid approach for extraction and normalization of temporal expression from English text document and it was evaluated on 4 datasets with good performance. The system can be used in applications like exploring search results on timeline, question answering, document summarization etc. In our work, we have taken document creation date as a reference date, but in some cases, it may depend on previous sentence. Further scope of improvement is to use heuristics around the text and previous sentences to select dynamic reference date.

References:

[1] G.Wilson, I.Mani, B.Sundheim, and L.Ferro. A multilingual approach to annotating and extracting temporal information. In Proceeding of workshop on temporal and spatial Information Processing Vol.13 Page 1-7.

[2] Mani, I. and Wilson, G. Robust Temporal Processing of News. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (Hong Kong, 2000), 69-76.

[3] H.Llorens, E.saquete and B.Navarro: TIPSEM(English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2.

[4] Oleksandr Kolomiyets, Marie-Francine Moens, KUL: Recognition and Normalisation of temporal expressions :Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 325–328, Uppsala, Sweden, 15-16 July 2010.

[5] M. Negri and L. Marseglia. Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004. Technical ReportWP3.7, Information Society Technologies, February 2005.

[6] Jannik Strotgen, Michael Gertz HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions:Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 321– 324,Uppsala, Sweden, 15-16 July 2010.

[7]Jordi Poveda, Mihai Surdeanu and Jordi Turmo.A comparison of Statistical and Rule Induction Learners for Automatic Tagging of Time Expressions in English

[8]Jelena Jacimovic: Recognition and Normalization of Temporal Expressions in Serbian Texts: *BCI'12,* September 16–20, 2012, Novi Sad, Serbia.

[9] Angel X.Chang, Christhpher D.Manning SUTime: A Library for Recognizing and Normalizing Time Expressions.

[10] WikiWars: A New Corpus for Research on Temporal Expressions: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing MIT, Massachusetts, USA, 9-11 October 2010.

[11] J. Lafferty, F. Pereira, and A. McCallum. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning, 2001.

[12] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In Proceedings of Human Language Technology-NAACL, 2003.

[13] A. McCallum and W. Li. Early results for Named Entity Recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the 7th CoNLL, 2003.

[14] W. Cohen. Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data, 2004.