Automatic TimeLine Generation of News Events

Dushyanta Dhyani

TECHNISCHE UNIVERSITÄT DARMSTADT

Internship Talk

Supervisors:

Prof. Dr. Iryna Gurevych
Nils Reimers

Project Duration:

1/1/2015-30/6/2015

Overview

- Motivation
- Goals
- Definitions
- Related Work
- Data Description
- Proposed Approach
- Features
- Results
- Discussion
- Contribution
- Future Work
- References

Motivation

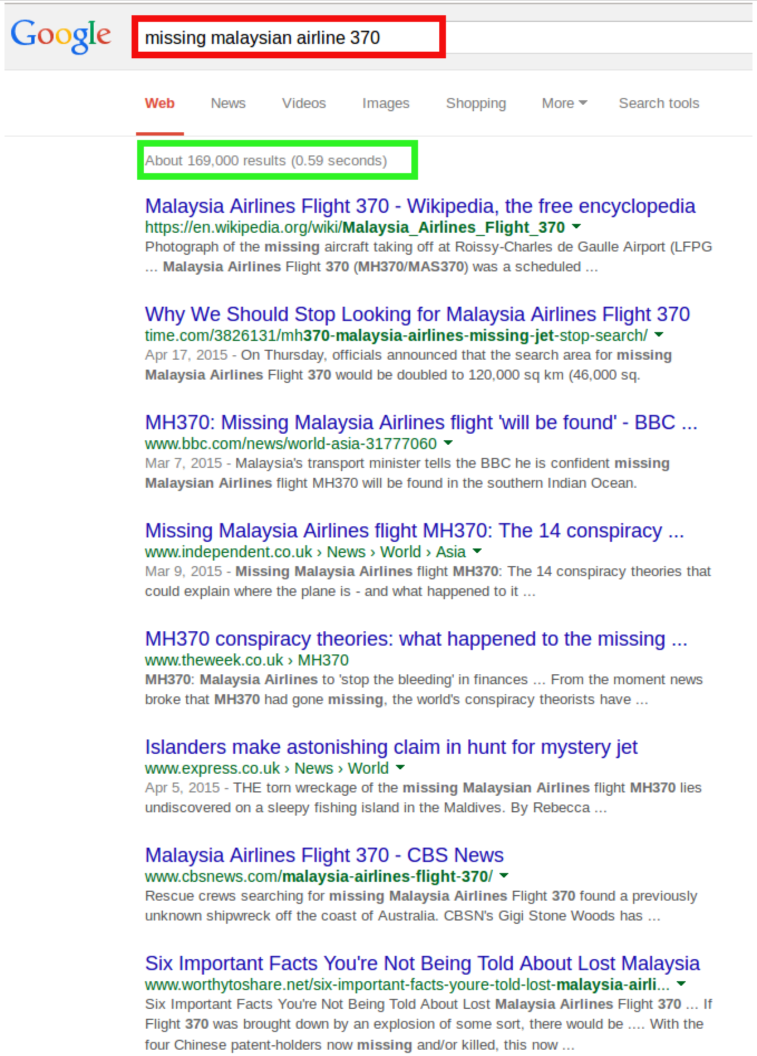
TECHNISCHE UNIVERSITÄT DARMSTADT

What we want?

Information about an event with respect to its evolution.

What we get?

Results are ordered based on their relevance to the search query.



Motivation

What we would prefer?

March 8,2014 - Kuala Lumpur-Beijing airliner with 239 people on board disappeared

↓

March 24,2014 - Malaysian Prime Minister Najib Razak made official statement after receiving reports from satellite data from Inmarsat, a UK Company

↓

Since 30 March 2014, the search is now being coordinated by the Joint Agency Coordination Centre (JACC), an Australian government agency

↓

January 29,2015 - Director General of the Department of Civil Aviation Malaysia, Azharuddin Abdul Rahman, announced that the status of Flight 370 would be changed to an "accident"

Goals - Stage 1

Extract **Relevant Information** from articles. This includes:

a) **Who** is involved? (**Persons/Organizations**)

b) **Where** did it happen? (**Location**)

c) **When** did it happen? (**Time**)

d) **What** happened? (**Events**)

Goals – Stage 2 & 3

Linking of relevant information snippets. This includes:

a) **Intra-Document Linking** - Who did what when and where?

b) **Inter-Document Linking** – Relation among events,participants,etc. Across a set of documents.

Use-Case Based Filtering

a) Given a **user defined information need**, we need a method to exploit the created ontology from Stage 2.

Definitions - Event

Something that happens or a situation that occurs.

It consists of four components

a) Action – **What** happens in the event.

b) Participants – **Who** or **What** is involved

c) Time – **When** the event happens

d) Location – **Where** the event happens

[Source : Guideline for ECB+ Annotation of Events and their Coreference [1]]

Definitions - Event

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: On Monday Lindsay Lohan checked into rehab in Malibu, California after car crash.

1.Action	Checked into;Crash	
2. Time	On Monday	
3. Location	Rehab in Malibu, California	
4. Participant	Human	Lindsay Lohan
	Non-Human	Car

Definitions – Human-Participants

Event Participants of entity type **PER,ORG** but also metonymically used **GPE, FAC and VEH** when **referring to population or government.**

a) Human_Part Per – Refers person entities and is limited to humans.
e.g. The **President** of the U.S.
e.g. The **family**

b) Human_Part_Org – Organization entities limited to corporations, agencies and other groups of people.
e.g. **Navy** attacked four LTTE boats.

c) Human_Part_GPE – Geo-political entities i.e. geographical regions defined by **political and/or social groups** referring to a **population or a government.**
e.g. **Poland** and the **US** signed a $34 million deal.

Definitions – Human-Participants

d) Human_Part_Fac – Facility Entities i.e. buildings and other permanent manmade structures and real estate improvements **referring to people using or managing them**.
e.g. The **school** decided to find a new location.

e) Human_Part_Veh – Vehicle Entities used in reference to a population or a government usually occuring with geo adjectives.
e.g. U.S. **ships** attacked 3 Iraqi patrol boats.
e.g. Serbian **tanks** attacked Croatian cities.

f) Human_Part_Met – Metonymically expressed Human Participants of events.
e.g. 30% of the **households**.
e.g. The **crown** gave its approval.

g) Human_Part_Generic - Generic mentions referring to a class or kind of human participants without pointing to any specific individual or individuals of a class.
e.g. **One** should treat others.
e.g. She loves working with **kids**.

Definitions – Non Human-Participants

NON_HUMAN_PART which is meant for **ALL remaining entity mentions** i.e. besides human participants of events, event times and locations – that contribute to the meaning of an event action.

Example
a) sharpen a **pencil** with a **knife**.
b) I hate **Mondays**.

Task-Event Extraction

Event Mentions are usually **noun phrases** or **verb phrases** that clearly describe events.

- It might be expected that event actions could be extracted reasonably well by identifying verb groups and event arguments.
- For instance we could apply SRL techniques to identify the *Agent* and *Patient* of each *predicate*. However, most SRL systems capture only *verb predicates.*
- Thus they would miss event mentions described via *noun phrases*.
- e.g. WWDC (World Wide Developers Conference)

Related Work

- EVITA System used a Bayesian Approach combined with Wordnet info.[2]
- Cybulska et al. Used Rule based event classification system using Historical texts.[3]
- ClearTK-TimeMl – Bethard et al. Used structural and syntactic features with an ensemble of SVM and Max Ent. Classifiers.[4]
- ATT(1|2|3) – Jung et al. used a Binary Max Ent Classifier using lexical, syntactic and semantic features.[5]
- TipSem – Used CRF's emphasizing on Semantic Roles.[6]

Data Description

ECB+ Corpus

Number of Documents : 982

Number of Topics : 43

Number of Sentences : 17075

EVENT/ACTION ANNOTATIONS

TOKEN	***COUNT***
I-ACTION	**2005**
B-ACTION	**11406**
O	**363951**

Data Description

TECHNISCHE UNIVERSITÄT DARMSTADT

ECB+ Corpus

PARTICIPANT ANNOTATIONS

ANNOTATION	COUNT	ANNOTATION	COUNT
I-PER	9186	B-PER	134
I-ORG	3793	B-ORG	23
I-PART	4457	B-PART	98
I-VEH	106	B-VEH	1
I-GPE	158	B-GPE	3
I-GENERIC	563	B-GENERIC	2
I-FAC	8		
I-MET	171		
O	358662		

Data Description

TimeBank

Training Data

Number of Documents	162
Number of Sentences	2176
Number of Tokens	53450
Number of Event Mentions	5688

Testing Data

Number of Documents	20
Number of Sentences	413
Number of Tokens	9613
Number of Event Mentions	968

Data Description

Inter Annotator Agreement

ECB+ Corpus

Category	Score
Actions/Events	81.1%
Participant	87.7%

TimeBank

Category	Score
Event	87%

Data Description

TECHNISCHE UNIVERSITÄT DARMSTADT

ECB+ Corpus – Issues/Shortcomings

a) Only **1840 (10%)** sentences are completely annotated.

Annotation	Original	New
I-ACTION	2005	1080
B-ACTION	11406	6832
O	363951	34898

Data Description

TECHNISCHE UNIVERSITÄT DARMSTADT

ECB+ Corpus – Issues/Shortcomings

- **a)** Only **1840 (10%)** sentences are completely annotated.

ANNOTATION	ORIGINAL	NEW	ANNOTATION	ORIGINAL	NEW
I-PER	9186	4503	B-PER	134	65
I-ORG	3793	1970	B-ORG	23	10
I-PART	4457	2067	B-PART	98	55
I-VEH	106	63	B-VEH	1	1
I-GPE	158	104	B-GPE	3	3
I-GENERIC	563	245	B-GENERIC	2	0
I-FAC	8	0			
I-MET	171	96			
O	358662	33628			

Data Description

ECB+ Corpus – Issues/Shortcomings

a) Only **1840 (10%)** sentences are completely annotated.

b) The News Documents **do not contain the date** of the news article which is presumably required by all the Temporal annotation tools.

c) Inconsistent annotation

ECB+ Corpus – Issues/Shortcomings

c) Inconsistent annotation

Example -

I)

Police	have	arrested	a	man
I-PER	O	O	O	I-PER

Police	arrest	allegedly	drunk	driver.
O	O	O	O	I-PER

II)

the	**next**	**generation**	macbook	pro
O	I-PART	I-PART	I-PART	I-PART

a	**new**	macbook	pro
O	O	I-PART	I-PART

III)

in	**his**	left	knee
O	O	O	I-PART

Data Description

TimeBank – Issues/Shortcomings

a) Does not have Participant annotations.

b) Does not have data segregation based on Topics, so difficult to analyze the Cross Domain adaptation of the Model developed.

Proposed Approach – Sequence Tagging

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Issues with conventional Classification techniques in the current context?**

  Ignores the latent sequential properties of text where correlation of neighbouring labels is important.

- **Reliability?**

Sequence tagging solutions have proven to provide state of the results in tasks like NER, POS Tagging.

- **Strategy**

  Use CRF's with various features extracted from a given window size

Features

- **Covered Text / Tokens**
- **Case -**
  Numeric/ All Lower / All Upper / Initial Upper / Other
- **Lemma**
- **Word Shape**
- **Word Prefix** -
  Length 1 to 5
- **Word Suffix** -
  Length 1 to 5
- **Part of Speech Tags**
- **Named Entity Tags**

Features

- **Dependency Parsing Features**
  a) Dependency Type
  b) Dependency Governor
  c) Dependency Governor Part of Speech

- **Semantic Role Labeling Features**
  a) Semantic Governor
  b) Semantic Governor Role -
  play.01, play.02, etc.
  c) Semantic Argument Role-
  The role of the predicate that governs the longest phrase of which the argument is a part.

Features

Word Embeddings

Used the Pre-trained Vector space Model from Google News Corpus.
Simple performed a K-means clustering with the parameters

K= 1000
No. of Iterations = 200

Results – Event/Action Extraction

ECB+ Corpus

Chunk Based

Category	Precision	Recall	F1
Action	79.02%	67.88%	73.02%

Results – Participant Extraction

ECB+ Corpus

Chunk Based

Category	Precision	Recall	F1
Org	66.5%	37.11%	47.64%
Part	67.39%	9.34%	16.40%
Per	80.49%	60.06%	68.79%
Weighted Macro Average	76.91%	44.66%	56.51%

Results – Event/Action Extraction

TECHNISCHE UNIVERSITÄT DARMSTADT

TimeBank – Comparison with Other Methods (Recognition Only)

Category	Precision	Recall	F1
Our Approach	78.93%	**82.73%**	80.78%
TipSem	**83.51%**	82.28%	82.89%
ClearTK	81.4%	76.38%	78.81%
ATT	81.44%	80.67%	81.05%
KUL	80.69%	77.99%	79.32%
FSS	63.13%	67.11%	65.06%

Results – Feature Significance (Event Extraction in ECB+ Corpus)

Basic Model – Included Features
Case,Lemma,Token,WordShape,Prefix,Suffix

Extended Model – Included only feature from
Word Embeddings, Dependency Parsing, NER , POS , SRL

Model	Score
Basic	66.82%
Basic+NER	67.76%
Basic+SRL	68.64%
Basic+Embeddings	69.02%
Basic+Dependency	69.19%
Basic+POS	71.20%

Result – Data Dependence

Discussion

TECHNISCHE UNIVERSITÄT DARMSTADT

Cases of Failure in ECB+ (Event Detection)

Token	Gold	Predicted
the **parking** lot crosswalk	O	I-ACTION
weighing in at 6.6 pounds	O	I-ACTION
next **generation** of Macbook Pro	O	I-ACTION
Apple unveils new macbook at **Wwdc**	I-ACTION	O
Global **marketing** vp phil schiller	O	I-ACTION
Macworld expo 2009	I-ACTION	O

Discussion

- **Cases of Failure in ECB+ (Participant)**

Token	Gold	Predicted
Police have arrested	I-PER	I-ORG
Apple has finally brought	I-ORG	I-PER
Sam's club in bloomington	O	I-ORG
miss the **team** 's final two games	O	I-ORG

Contributions

- Analyzed the suitability of ECB+ Corpus for the task

- Created a robust Base Model as a foundation for subsequent research that performs well even without involving complex features.

- Identified attributes that needed more investigation.

Future Work

- Add Temporal Annotation Components for Chronological Sequencing of Events . (Currently in Progress)
- Add Co-reference resolution for identifying relation among various entities.
- Identifying the relative significance of various Events / Participants.

References

1. Cybulska, Agata, and Piek Vossen. Guidelines for ecb+ annotation of events and their coreference. Technical report, Technical Report NWR-2014-1, VU University Amsterdam, 2014.
2. Saurí, Roser, et al. "Evita: a robust event recognizer for QA systems." Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005.
3. Cybulska, Agata, and Piek Vossen. "Historical event extraction from text." Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics, 2011.
4. Bethard, Steven. "ClearTK-TimeML: A minimalist approach to TempEval 2013." Second Joint Conference on Lexical and Computational Semantics (* SEM). Vol. 2. 2013.
5. Jung, Hyuckchul, and Amanda Stent. "ATT1: Temporal annotation using big windows and rich syntactic and semantic features." Second Joint Conference on Lexical and Computational Semantics (* SEM). Vol. 2. 2013.
6. Llorens, Hector, Estela Saquete, and Borja Navarro. "Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2." Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2010.
7. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

Questions/Suggestions?

TECHNISCHE UNIVERSITÄT DARMSTADT

UKP