Named Entity Patterns across News Domains

Masinizah Mohd
Department of Computer and Information Sciences
University of Strathclyde, Livingstone Tower, 26 Richmond Street,
Glasgow, G1 1XH, United Kingdom
masnizah@cis.strath.ac.uk

Abstract: A new event tracking approach is proposed based on the identification of named entity (NE) patterns such as Who, What, Where and When, and their relationship with news domains such as Politics, Economy, Government and Entertainment. This research comprises three parts. The first part uses a set of user studies to identify NE patterns and their relationship with news domains. Second part is to design a prototype system based on NE patterns. The final part evaluates the prototype event tracking system. This paper described the first part which is to evaluate the importance of NE across news domains. We have achieved a better understanding on NE patterns by identifying the distribution of NE across news domains.

Keywords: information retrieval, event tracking, named entity

1. INTRODUCTION

Topic Detection and Tracking (TDT) applies to the detection and tracking of events from a stream of stories. This stream may or may not be pre-segmented into stories, and the events may or may not be known to the system, where the system may or may not be trained to recognize specific events. This leads to the identification of three technical tasks to be addressed in the present TDT study. These are, namely, the tracking of known events, the detection of unknown events, and the segmentation of a news source into stories.

The tracking task associates incoming stories with events known to the system. An event is defined by its association with stories that discuss that particular event. Thus each target event is characterized by a list of stories which discuss it. Each successive story must be classified according to whether or not it discusses the target event in the tracking task. Therefore, the study corpus is divided into two parts, with the first part being the training set and the second part being the test set. Each of the stories in the training set is flagged as to whether it discusses the target event or not, and these flags (and the associated text of the stories) are the only information used for training the system to classify correctly the target event.

The TDT tracking task is fundamentally similar to an information filtering task [9]. Each begins with a representation of a topic and then monitors a stream of documents, making decisions as they arrive. Each document is assigned a score for that topic and, if the score is high enough, it is retrieved. Filtering simulates interacting with the user to supervise the process, whereas tracking operates as if the user were not there. Systems may be adaptive in that they “guess” that a story is on topic, but they do not receive human confirmation that they were correct.

Recent research in TDT has investigated named entities rather than words because TDT investigates the organization of information by event rather than by subject [2, 4, 5, 6, 7, 8]. This short paper has identified NE patterns across news domains so that we can improve the event detection and tracking.

2. EVENT TRACKING BY USING NAMED ENTITY APPROACH

In this section, we review the literature concerning the usefulness of using named entity (NE) patterns across domains to improve the performance of an event tracking system. As mentioned above, tracking is basically a simpler version of the classic IR filtering task. Thus, an increasing number of information retrieval and machine learning techniques have been applied. These include k-Nearest Neighbour (kNN) classification, Decision Tree induction, a variety of Language Modelling (LM) approaches and relevance based filtering methods [1, 10], and systematic analysis of their behaviour on event tracking has only just begun.

A TDT system draws a distinction between events and topics. As Yang et al. [10] note, “The USAir-427 crash is an event but not a topic, and airplane accidents is a topic but not an event”. Studies by various authors [10, 3, 7] have emphasized the importance of having a better understanding and a clear definition of an event. Indeed, it appears that event is a term which easy to understand at the intuitive level but hard to define precisely. An event comprises at the very least what happened, where it happened, when it happened, and who was involved. In the Information Extraction field, within Message Understanding Conferences (MUC) has presented an approach to define an event using a Scenario Template (ST) task. The main goal of ST is to extract pre-specified event information and relate the event information to particular organisation, person, or other entities involved in the event [1].
The identification and extraction of NE word patterns, relate to the questions about Date (when), Location (where), Person (who) and Organisation (what/who) is crucial in TDT because detection of named entity allow us to characterize and detect event in document. This is particularly true for event topics. Our motivation is to have a clear understanding on NE patterns in event tracking.

We concluded that event is defined as a unique circumstances or a condition that involved the integration of the W’s elements as shown in Table 1 below:

| NE   | Description                                           | NE example                                                                 |
|------|-------------------------------------------------------|-----------------------------------------------------------------------------|
| Who  | This element refers to the actor or person that takes part in the event. | Alex Salmond won the Gordon constituency in the 2007 Scottish Parliament election. |
| Where| This element refers to the places or location that the event takes place.  | Scotland was the location for the 2007 Scottish Parliament election.         |
| When | This element refers to the date or time that the event takes place.      | 3rd May 2007 was the election day.                                           |
| What | This element refers to the subject, occasion, body or activity that involved in the event. | The SNP won the 2007 Scottish Parliament election.                             |

TABLE 1: Description of NE showing examples in italics

Research approaches in TDT have, in common, the use of NE at the sentence level [2, 4, 5, 6, 8]. Collections in TREC 2003 and TREC 2004 novelty track are classified into opinion topics and event topics. We note that the opinion patterns in sentences are used for improving the performance of novelty detection for opinion topics while NE plays a more important role for event topics. However, simple uses of NE seems to be not very helpful for improving the performance of TDT system, therefore, the challenging questions are how to effectively make use of this NE, and what kinds of additional and critical information patterns will be effective.

Kumaran and Allan [4] observed that certain categories of news are better tackled using solely NEs such as Elections, Accidents, Violence and War, New Laws, Sports News, and Political and Diplomatic Meetings. While for other news categories, it was better to use topic terms. Thus, it is important for this pilot study to experiment with other general news domains and this strongly motivates the work on NE patterns and their relationship with the news domain which may contain important and relevant information. In addition, there have been no studies to date which focus on the distribution pattern of NE across news domains.

3. RESEARCH GOALS AND OBJECTIVES

Our overall research programme will consider the constraints and issues in the tracking task and come out with a good event tracking system. The specific objectives of this research are to:

1. Identify NE patterns across news domains;
2. Design a prototype of an interactive and adaptive event tracking system based on NE patterns;
3. Experiment and to evaluate the system performance, in terms of precision at top ranks.

This paper tackles objective 1 in order to have a better understanding of NE patterns by identifying the distribution of NE across news domains.

4. PILOT STUDY

A pilot study was undertaken to support the first objective of the research. It consisted of two phases; the first was conducted in early February 2007, while the second took place in early March 2007. 10 postgraduate students from the Scottish Centre for Journalism Studies (SCJS), University of Strathclyde were selected in this study.

4.1 Objectives

The objectives of this pilot study were to:

a. Identify NE patterns (Who, Where, When and What) across news domains (Politics, Economy, Government and Entertainment);
b. Measure the importance level of NE across news domains;
c. Measure the level of agreement in the keywords given by the respondents.

These objectives will guide us to design the prototype system by investigating NE patterns from the frequency and the importance aspect. That is, are some types of NE more important and more frequent in some domains?

4.2 Instruction and online questionnaire
An online questionnaire was placed on a server in the department of Computer and Information Sciences (CIS) at University of Strathclyde with the link as https://devweb2006.cis.strath.ac.uk/~masnizah/cgi-bin/survey/index.htm. It was designed using HTML and CGI scripting as a front end to receive the data via email. Respondents were given the instructions including the description of named entities, as described in Table 1, and the privacy statement at the main page. They were asked to read each document. They were then required to give the important keywords that best describe the document and tick their importance on a scale of 1 to 5, with 1 being the least important and 5 being most important. Then the respondents had to classify the keywords according to which type of named entity is being mentioned. Finally at the end of the online questionnaire, respondents were also given an open-ended question as an opportunity to supply any comments or suggestions regarding the pilot study.

4.3 Corpus and distribution
We constructed a corpus of 40 documents chosen from the CNN News, Associated Press and Scotsman.com News. Each domain consisted of 10 documents. The sources and documents were current stories related to Scotland or Glasgow as all respondents were living in Glasgow such as documents P5 and P10 in Politics; E1, E4 and E8 in Economy; G1 in Government; and Et3 in Entertainment. The documents are shown in Table 2.

| Domain | Doc.ID | Document |
|--------|--------|----------|
| Politics | P1 | Saddam verdict timing ‘suspect’ |
| | P2 | Protests as Bush visits Indonesia |
| | P3 | Blair meets troops in Afghanistan |
| | P4 | Support for Labour falls to 20-year low |
| | P5 | Nationalists say figures prove Scots subsiding the UK |
| | P6 | Mubarak warns on Saddam execution |
| | P7 | Bush offers Veterans Day tribute |
| | P8 | War on terror ‘could last 30 years’ |
| | P9 | SNP ‘wants to lay down law on jail sentences’ |
| | P10 | Gordon Brown has ‘sold his Scottish soul’ says SNP chief |
| Economy | E1 | Scottish Water customers ‘saving £90 a year on bills’ |
| | E2 | Rainy day for Amazon as profits drain away to £10m |
| | E3 | Jobless totals rise over last year |
| | E4 | Skills gap seen as biggest threat to Scottish firms |
| | E5 | Britain branded ‘card fraud capital’ |
| | E6 | EU to call for more labour-law flexibility |
| | E7 | Proof of Concept scheme will generate £125 million |
| | E8 | Scottish house price rises soar above rest of UK |
| | E9 | Finance sector now 10% of UK economy |
| | E10 | Economic impact of immigrants in the spotlight |
| Government | G1 | Students reject union membership |
| | G2 | Poorest families may get cash for fuel |
| | G3 | Extra money for mentors to tackle bullying |
| | G4 | Behaviour measures target parents |
| | G5 | PM unveils plans for ‘supernannies’ |
| | G6 | Latest figures on migrants released |
| | G7 | Dutch Plan To Ban Burqas And Masks |
| | G8 | Low income families grants call |
| | G9 | NHS ‘needs time to balance books’ |
| | G10 | Bourne seeks Welsh spending boost |
| Entertainment | Et1 | Cruise and Holmes leave Italy after wedding |
| | Et2 | Madonna ‘wanted to adopt baby girl’ |
| | Et3 | Louis: Macs Can Beat The Axe |
| | Et4 | Virtual “Big Brother” to be launched in Second Life |
| | Et5 | Ramsay gets his hands on Emmy award |
| | Et6 | Tom and Katie ‘had already married’ |
| | Et7 | Madonna rubbishes new adoption reports |
| | Et8 | Penguins beat Bond at US cinemas |
| | Et9 | X Factor’s Robert upbeat on future |
| | Et10 | OJ Interview Cancelled: Family’s Relief |

TABLE 2: Documents across domains

A total number of 12 documents was given to each respondent from this pilot study. The documents’ distribution was based on a repeated Latin square. The reason for this was to have a balanced distribution of the documents to every respondent such that each respondent receives documents from the four domains and every document will be viewed by 3 respondents. As a result, this distribution allow the comparison of the keywords from the same
document and support the third objective of the pilot study which is to measure the level of agreement among respondents.

4.4 Results and discussion
A total of 557 data points or keywords from the 40 documents was analysed and to identify:

1. NE patterns across news domains;
2. The importance level of NE and news domains;
3. The level of agreement in the keywords and news domains.

4.4.1 NE patterns across news domains
Table 3 summarizes the distribution of NE across domains. Politics has more What NE (40.8%) and Who NE (36.3%). Economy has more What NE (64.1%) and Who NE (18.8%), Government has more What NE (54.0%) and Who NE (36.3%) and, Entertainment has more What NE (45.2%) and Who NE (37.8%). The most striking result to emerge from the data is that the dominant type of named entity was the What; its percentage of its occurrence across domains accounts for at least 40% of keywords. The highest percentage of What NE was in Economy (64.1%). However, the distribution of When NE within domains was less than 6% and surprisingly the percentage of it was 0% in Government.

| DOMAINS      | N | Named entity |
|--------------|---|--------------|
| Politics     | 157| 57 (36.3%) | 28 (17.8%) | 8 (5.1%) | 64 (40.8%) |
| Economy      | 128| 24 (18.7%) | 20 (15.6%) | 2 (1.6%) | 82 (64.1%) |
| Government   | 137| 50 (36.5%) | 13 (9.5%)  | 0 (0%)   | 74 (54.0%) |
| Entertainment | 135| 51 (37.8%) | 21 (15.5%) | 2 (1.5%) | 61 (45.2%) |

TABLE 3: Distribution of NE across domains

The calculated ranges of NE across domains was 40%–65% for What, 18%–38% for Who, 15%-18% for Where and 1%-6% for When. This result is interesting as the Who (32.7%) and What (50.4%) are perceived as being more frequent to be chosen as the type of named entity in this study. What and Who are the top NE across domains. The sequence of NE distribution was identical across news domains, with What more than Who NE followed by Where NE and the least was When NE. However from the Kruskal-Wallis Test, there is a statistically significant difference in the distribution of NE frequency within domains, $x^2(3) = 17.1$ and $p = 0.001$. A Chi-square test was performed and there is strong evidence to indicate a relationship between NE and news domains was found, $x^2(9, N = 557) = 32.1$ and $p < 0.001$. This shows that the distribution pattern on the type of NE is domain dependent across the Politics, Economy, Government and Entertainment.

4.4.2 The importance level of NE and news domains
Table 4 summarizes the distribution of the importance level of NE across domains. Findings showed that for the Very Important level of NE, What was the highest occurrence of NE across domains and the highest percentage of What NE was in Economy (63.4%). While for the Important level of NE, again, What was the highest occurrence of NE in Economy, Government and Entertainment except in Politics, where Who NE has a higher percentage (36.4%). For the Fairly Important level of NE, What was the highest occurrence of NE in Politics, Economy and Government except in Entertainment, where Who NE has a higher percentage (44.0%). Surprisingly, for the Quite Important level of NE, Where has the highest percentage in Politics (44%) and Entertainment (100%), What was the highest occurrence of NE in Economy (100%) and Who; What NE shared the same percentage (50%) in Government.

We used the Spearman’s rank correlation test to predict the correlation between the importance level of NE and news domains. The findings indicated that there is no significant relationship between the importance level of NE and news domains where $p > 0.001$. The importance level is domain independent and it shows that respondents are giving different weightage on NE. Although What is the top NE but it is not necessarily Very Important across news domains.
4.1.3 Level of agreement in the keywords and news domains

Table 5 shows the overlap values that ranged from 0.11 to 1.00. Four documents with perfect agreement (1.0) were "NHS needs time to balance books" (G9), "Cruise and Holmes leave Italy after wedding" (E11), "Madonna wanted to adopt baby girl" (E12) and "X Factor's Robert upbeat on future" (E9). 2 documents with the lowest overlap value (0.11) were "Support for Labour falls to 20-year low" (P4) and "Bourne seeks Welsh spending boost" (G10). The findings revealed that the Entertainment has the highest value of mean overlap (0.67) compared to the Politics (0.42), Economy (0.41) and Government (0.47) of mean overlap. Finding also found the disagreement in the type of named entity given to the same keyword such as in the article P9 and P2. Respondents have classified the keyword "Holyrood election as When and What, and the keyword "Terrorist" as "What NE and Who NE. The reason for this is not clear but it may have relation with respondent's interest and familiarity with the genre of documents. The overlap value from this study showed the variation in the individual judgements of the keyword from the document. Thus, this will guide us for the selection of documents and to measure the system performance by the keywords given.

4.1.3 Level of agreement in the keywords and news domains

The level of agreement in the keywords given by the respondents was calculated by using the overlap value [9] which is the intersection of keywords divided by the union of keywords from the same document. Thus, an overlap of 1.0 means perfect agreement and an overlap of 0.0 means none of the respondents agreed with the keywords given. We used stemming to identify the intersection of keywords such as launched, launching and launch were counted as a single term. We also conflated different keywords referring to the same context and meaning, such as Tony Blair and Prime Minister (PM), Nationalists and SNP, Hamid Karzai and President Hamid Karzai, unemployment and jobless, Labour UK and Labour, Saddam Hussein and Saddam, British troops and troops, Hosni Mubarak and Mubarak, Louis Walsh and Louis.

### TABLE 4: Importance level of named entity across domains

| DOMAINS       | N   | NE (%) | Who | Where | When | What |
|---------------|-----|--------|-----|-------|------|------|
| Politics      | 77  | 30 (39.0%) | 8 (10.4%) | 5 (6.5%) | 34 (44.2%) |
| Economy       | 71  | 17 (23.9%) | 8 (11.3%) | 1 (1.4%) | 45 (63.4%) |
| Government    | 68  | 30 (44.1%) | 2 (2.9%)  | 0 (0%)   | 36 (52.5%) |
| Entertainment | 78  | 32 (41.0%) | 5 (6.4%)  | 1 (1.3%) | 40 (51.3%) |

| ID | Overlap | ID | Overlap | ID | Overlap | ID | Overlap |
|----|---------|----|---------|----|---------|----|---------|
| P1 | 0.38    | E1 | 0.60    | G1 | 0.57    | E1 | 1.00    |
| P2 | 0.50    | E2 | 0.40    | G2 | 0.50    | E2 | 1.00    |
| P3 | 0.57    | E3 | 0.57    | G3 | 0.33    | E3 | 0.80    |
| P4 | 0.11    | E4 | 0.29    | G4 | 0.80    | E4 | 0.60    |
| P5 | 0.83    | E5 | 0.67    | G5 | 0.67    | E5 | 0.67    |
| P6 | 0.33    | E6 | 0.60    | G6 | 0.33    | E6 | 0.80    |
| P7 | 0.67    | E7 | 0.33    | G7 | 0.33    | E7 | 0.67    |
| P8 | 0.33    | E8 | 0.40    | G8 | 0.33    | E8 | 0.43    |
| P9 | 0.38    | E9 | 0.57    | G9 | 1.00    | E9 | 1.00    |
| P10| 0.50    | E10| 0.67    | G10| 0.11   | E10| 0.43    |

### TABLE 5: Overlap values for documents across domains

Table 5 shows the overlap values that ranged from 0.11 to 1.00. Four documents with perfect agreement (1.0) were "NHS needs time to balance books" (G9), "Cruise and Holmes leave Italy after wedding" (E11), "Madonna wanted to adopt baby girl" (E12) and "X Factor's Robert upbeat on future" (E9). 2 documents with the lowest overlap value (0.11) were "Support for Labour falls to 20-year low" (P4) and "Bourne seeks Welsh spending boost" (G10). The findings revealed that the Entertainment has the highest value of mean overlap (0.67) compared to the Politics (0.42), Economy (0.41) and Government (0.47) of mean overlap. Finding also found the disagreement in the type of named entity given to the same keyword such as in the article P9 and P2. Respondents have classified the keyword "Holyrood election as When and What, and the keyword "Terrorist" as "What NE and Who NE. The reason for this is not clear but it may have relation with respondent’s interest and familiarity with the genre of documents. The overlap value from this study showed the variation in the individual judgements of the keyword from the document. Thus, this will guide us for the selection of documents and to measure the system performance by the keywords given.

5. CONCLUSION

BCS IRSG Symposium: Future Directions in Information Access (FDIA 2007)
Recent research in TDT has used information on NE and this requires more effort to understand all the NE. Thus, our study has investigated the type and the use of NE in news domains. We have achieved a better understanding on the correlation of the type of NE across domains has been achieved. This study revealed that there is a significant difference in the distribution of NE within domains. What is the most dominant type of named entity followed by the Who named entity, across news domains. This is interesting since the results give clues on what are the named entities that user would expect more to occur within domains.

We found that there is no significant relationship between the importance level of NE and news domains, indicating that the importance level is domain independent. This revealed the respondents perception on how often the named entity occurred and how important the named entity across domains. When has 0% of occurrence in Government but it can be Very Important type of named entity.

While the level of agreement in the keywords among the respondents is likely to be higher in Entertainment compare to other domains. The measure of agreement based on the overlap value is substantial to give a better understanding on the documents selection and how to use the keywords given to measure the system performance.

The findings have key implications for developing a prototype of a system design that should adapt on the user profile and enhance the understanding of the user needs. Further research is needed to investigate the specific news to see the importance of When and Where NE.

Acknowledgements: I would like to thank Fabio Crestani and Ian Ruthven for their help and advice, the postgraduate students (SCJS) who made the pilot study possible and Richard Bache for advice on drafting the paper.

REFERENCES

[1]. Allan, J. (2002) Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers, Norwell, MA.

[2]. Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R. and Caputo, D. (1999) Topic-based novelty detection. Summer workshop at CLSP final report. Baltimore, MD. August.

[3]. Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K. and Liberman. M. (2002) Corpora for topic detection and tracking. In Allan, J. (ed), Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers, Norwell, MA.

[4]. Kumaran, G. and Allan, J. (2004). Text classification and named entities for new event detection. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, 25–29 July, pp.~297–304. ACM.

[5]. Li, X. (2006) Sentence Level Information Patterns for Novelty Detection. Ph.D. University of Massachusetts at Amherst.

[6]. Li, X. and Croft, W.B. (2005) Novelty Detection Based on Sentence Level Information Patterns. Proceedings of ACM CIKM’05, Bremen, Germany, 31st October - 5th November, pp.~744–751. ACM.

[7]. Makkonen, J., Ahonen-Myka, H. and Salmenkivi, M. (2002) Applying semantic classes in event detection and tracking. Proceedings of the International Conference on Natural Language Processing (ICON’02), Mumbai, India, 18-21 December, pp.~175–183.

[8]. Otterbacher, J., Erkan, G. and Radev, D.R. (2005) Using Random Walks for Question focused Sentence Retrieval. Proceedings of Human Language Technology / Empirical Methods in Natural Language Processing (HLT-EMNLP’05), Vancouver, Canada, 6-8 October, pp.~915–922. ACM.

[9]. Voorhees, E.M. and Harman, D. (2000) Overview of the Eighth Text REtrieval Conference (TREC-8). Proceedings of the Eighth Text REtrieval Conference (TREC-8), Gaithersberg, Maryland, 17-19 November, pp.~1–24. NIST Special Publication.

[10]. Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B.T. and Liu, X. (1999) Learning approaches for detecting and tracking news events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 14(4), 32–43.