# Link Detection Based on Named Entity Keywords in Turkish News Corpus

Hamid Ahmadlouei, Hayri Sever
{hamid,sever}@hacettepe.edu.tr
Department of Computer Engineering
Hacettepe University
Ankara, Turkey

Erhan Mengusoglu
emengusoglu@thk.edu.tr
Computer Engineering
University of Turkish Aeronautical Association
Ankara, Turkey

***Abstract*— In this study, we investigate the influence of Named Entities (NEs) on the task of Story Link Detection (SLD), which is one of the important subtasks in Topic Detection and Tracking (TDT). TDT aims at developing algorithms for either clustering documents, e.g., online news, and then tracking new ones with respect to a predetermined topic or otherwise detecting a new topic. Furthermore, SLD focuses on determining whether the stories are about the same topic. Vector Space Model (VSM) was used as a base method in this work for "All-words" and Named Entities (NE) separately. We also investigated the effect of controlled entity intersection on the performance of previous VSM based methods. Combining these methods provided improvement in correctly estinating whether the stories are linked or not.**

***Keywords*— *story link detection; topic detection and tracking; vector space model; information retrieval; named entity.***

## I. INTRODUCTION

In recent years, there have been a growing number of online news sources. Having many options might be attractive for the user, but, on the other hand, the user might spend a substantial amount of time surfing the Internet in search for the needed information. If proper information retrieval (IR) techniques are used, helping the user reach the needed information becomes an easier task. So, in order to manage the huge increase in news articles, grouping similar articles and linking similar stories are indispensable steps for IR tasks.

An information retrieval system is composed of a corpus of documents, and access functions with capability of comparing the words of the query terms in user queries, and the terms of the documents in the corpus to determine the relevant documents. At this point, the basic function of information retrieval systems is to access all relevant documents in the corpus and comb out non-relevant ones in order to meet the information needs of users [1]. TDT is one of the most important tasks in the study of IR. Hence, recent academic studies on the Web IR systems mainly focus on the TDT program.

TDT studies aim to organize, identify and follow all kinds of stories published on the Web [2]. To accomplish this goal, TDT studies are divided into five main tasks: story segmentation, topic detection, topic tracking, first story detection and story link detection:

- Story Segmentation: determines story boundaries,

- Topic Detection: determines the subject of the story,

- Topic Tracking: follows a pre-determined story,

- First Story Detection: identifies stories not encountered previously,

- Story Link Detection: determines if two stories are linked or not,

SLD tasks are reported as the most important sub-tasks of TDT studies [3]. The purpose of SLD is to determine whether two independent stories are on the same subject or not [4]. "Story" in this paper is defined as a piece of news about a single "topic" in TDT problems. "Topic" is a seminal event or an activity along with all directly related events and activities. [2]. "Event" is a specific thing that happens at a specific time and place.

In this paper, we analyze story link detection and investigate word based and named entity based techniques. Our purpose is to detect whether two documents are linked or not. We present the performance of two techniques and show the improvement in the performance of a link detection system using a combination of these techniques. After performing the SLD task, the results of this sub-task can be applicable to other sub-tasks of TDT. In this respect, successful determination of whether two stories are on the same topic or not by using SLD, is expected to solve many problems for TDT [5].

This paper presents a combination of different techniques to improve the performance of link detection. A combination of methods in some cases provides an improvement in estimating whether two stories are linked or not. The work is evaluated on the Turkish news corpus, and the experimental results indicate that story link detection using a combination of methods can help obtain a better performance.

We used VSM as the base model in this work. Our experimental results indicate the result of VSM which is inspired by the co-sine similarity concept, is better than alternatives. Word-based (WB) and entity-based (EB) tests of this method are carried out separately.

We also control the intersection of named entities between two stories. Intersection checking is used in order to determine

 ISBN: 978-1-61208-457-2 

which two different news are on the same topic. Experiments are designed to assess how VSM performance is affected when entity intersection checking is used. We analyzed OR and AND logical combination of VSM with the Named Entity Intersection (NEI) method. Word-based and entity-based scenarios of VSM, OR/AND-logical combination with NEI are carried out separately.

Inspired by the study presented in [6], we defined a simple Named Entity Resemblance Function (NERF) in order to give more importance to the named entities between two stories. To enhance the effectiveness of named entities by simple normalization on naming entities between news articles, the similarity score between two news stories is calculated using the function in [6]. This method was not successful in Turkish news tasks, but it was successful in a Chinese study.

This paper is organized as follows. In Section 2, we give an outline of the related work within the topic. In Section 3, we talk about the methodology used in the paper. In Section 4, we give details of the tests carried out during our experiments and present their results. In Section 5, we present the conclusion by summarizing our contribution. In Section 6, we briefly note future studies that can be carried out as a follow-up for this work.

## II. RELATED WORK

Academic studies in the field of TDT story link detection task require identifying pairs of linked stories. In the story link detection systems which have been developed so far, the best technology for link detection relies on the use of cosine similarity between document terms vectors with Term Frequency - Inverse Document Frequency (TF-IDF) term weighting.

Some academic studies in the SLD field show that relevance models (RM) produce better results than other IR methods [7]. Some works on SLD based on VSM use the cosine similarity measurement between the data streams [8] [9] [10]. UMass has examined a number of similar measures in the link detection task, such as weighted sum and language modeling, and found that the cosine similarity produced the best results [10]. Additionally, a different study by the same authors showed that VSM performed better on the SLD task based on Turkish news [4].

Another point that information retrieval researchers focus on is how to select terms representing documents and weight them effectively. Document representation is an extremely important step in traditional IR systems as well as in TDT studies [11]. Depending on the study areas, word-based and entity-based methods are usually used for the representation of the documents [6] [12] [13].

In SLD task, many methods are used to compare the quantity of the overlapping words within two stories. Large numbers of overlapping words between two stories means there is a higher probability that they discuss the same topic. This approach formed the basis of all the methods that use vector space models [14].

Some studies on TDT task also claim that document representation using only words is not enough [15]. Experiments reported in [16] compared two different stories using named entities taking into account: person (who), location (where), time (when) and action (what) words. In a similar study, name, location and time information in the news are expressed in separate vectors and named entities such as name, place and time are extracted by automatic inference methods. Using the named entities to identify the most recent (newest) news provides a significant performance increase [6]. Researchers in both of those studies proposed similarity metrics based on intersection, especially while comparing the time and place [17].

The use of entity names on the Turkish corpus in order to improve SLD performance was not studied so far. Literature emphasizes that more in-depth studies should be done in this regard [18].

In some studies, a combination of different methods provides improvement for estimating whether two stories are linked. Furthermore, it is also reported in the literature that access performance is increased by using a combination of different methods [4] [6] [19].

Named entities are also used in determining news similarity in order to perform SLD task [18]. Similar studies for Turkish corpus, mainly by using machine-learning methods, that extract named entities (name, place, organization e.g.) automatically from texts are reported in [20] [21].

## III. METHODOLOGY

### A. TDT Test Collection

We used new event detection and topic tracking test collection (BilCol-2005), which was developed by the information retrieval group at Bilkent University. The Bilkent information retrieval group aims to develop effective and efficient information retrieval tools, with an emphasis on the Turkish language. The BilCol-2005 test collection comprises news stories from five different Turkish news sources on the Web (both broadcast news and daily news articles): CNN Turk, Haber 7, Milliyet, TRT, and Zaman. More information about BilCol-2005 is provided in [18].

In the BilCol-2005 corpus, 5,883 news stories were classified under 80 different topic titles, while the rest (203,442) has yet to be classified. In this study, tests were carried on the classified news stories.

Most of the studies on Named Entity Recognition (NER) subfield have been done for English, Chinese and Spanish texts. Studies for Turkish language texts have only started recently. Thereby, automatic NER methods in Turkish are still immature. So, in the preparation phase of the dataset, tagging of named entities was carried out manually. As in many IR systems, most studies focus on which words to select as named entities or keywords and how weighting of these keywords has to be done as well as how these weighted keyword will most effectively be compared [22] [23]. In this context, named entities in the dataset are tagged with the following labels: *"Person"*, *"Location"*, *"Organization"*, *"Time"*, *"Date"*, *"Percentage"*, *"Money"*, *"Unknown"*. The

 ISBN: 978-1-61208-457-2 

label "*unknown*" is used for tagging all entities in the text when tagging with the other labels above was not possible.

Here is an example of a labeled entity:

<Person>Shakespeare<Person>

<Location>Ankara<Location>

<Organization>Galatasaray<Organization>

<Date>1992<Date>

*B. Evaluation Methodology*

The performance is measured by obtaining *precision*, *recall* and *F-measure* values for each test. *Recall* is the proportion of retrieved relevant documents to total related documents and *precision* is the proportion of accessed relevant documents to total accessed documents. *F-measure* identifies the harmonic mean of precision and recall [24]. These three values are expressed mathematically in the following equations:

$$\Pr ecision = \frac{number-of-accessed-relevant-document}{number-of-accessed-document} \quad (1)$$

$$\mathrm{Re}call = \frac{number-of-accessed-relevant-document}{total-number-of-relevant-document} \quad (2)$$

$$F-Measure = \frac{2*\Pr ecision*\mathrm{Re}call}{\Pr ecision+\mathrm{Re}call} \quad (3)$$

In this paper, we assumed that high *precision* and high *recall* or higher *F-measure* values represent better results.

Studies on the combination of different methods generally increases the values of recall, yet, at the same time, retrieves a lot of unrelated documents, thereby decreasing precision values and degrading the overall system performance. Therefore, it is extremely important to develop combined models that would provide the best possible values for both precision and recall.

*C. SLD Methods Used In The Study*

Different types of documents may have different retrieval characteristics. Text retrieval methods are typically designed to find documents relevant to a query based on some criterion, such as cosine similarity. We used vector space model (VSM) as a base method for the Turkish corpus. The vector space model developed in the late 1960s is still a very popular approach and is commonly used in IR systems as a retrieval function [25]. Although this method has been widely used for SLD task of TDT studies, there is, to the best of our knowledge, no study that was carried out to apply it on a Turkish corpus [2][3].

VSM calculates the similarity between compared documents based on common term conflicts. So, we analyze the word-based and entity-based approaches in the first two steps of the experiments.

The steps used in our experiment are:

*1- VSM Word Based (WB)*

*2- VSM Entity Based (EB)*

*3- Named Entity Intersection (NEI)*

*4- VSM (WB) OR NEI*

*5- VSM (WB) AND NEI*

*6- VSM (EB) OR NEI*

*7- VSM (EB) OR NEI*

*8- Named Entity Resemblance Function (NERF)*

VSM is used with words and entity based approach as an access function. In information retrieval systems, which use this method, each document is shown as a vector of a collection of $t_1$, $t_2 \ldots t_n$ single words. Coefficient values of $t_1$, $t_2, \ldots t_n$, are determined based on the number of times that related word appears in the collection ($t_i$).

In traditional IR methods, a general approach for representing the vector coefficients is identified as the *idf-weighted cosine coefficient* and is shown as *tf.idf* (*term frequency*, *inverse document frequency*). Similarity between the two vectors (*a* and *b*) is calculated by applying Equation 1 where $tf_a\ (w)$ represents the frequency of word *w* in the document *a*, $tf_b\ (w)$ represents the frequency of word *w* in document *b*, and $idf\ (w)$ represents the frequency of word *w* in all documents in the corpus.

$$sim(a,b) = \frac{\sum_{w=1}^{n} tf_a(w).tf_b(w).idf(w)}{\sqrt{\sum_{w=1}^{n} tf_a^2(w)}.\sqrt{\sum_{w=1}^{n} tf_b^2(w)}} \quad (4)$$

In the second step, by using the determined entity names, we created entity vectors for each article document. By applying Equation 4, we can calculate the similarity between the entity vectors. With this method, the biggest challenge is that some documents may not have enough named entities for creating a good quality entity vector. If the vector created is very sparse, it will not be good enough to make comparisons.

To resolve this problem, in the third step, we controlled the named entity intersection between the stories. When two news stories are compared and even if only one entity intersection is found, then these two news stories were determined as linked. In this step, the entity intersection control is examined by determining whether these two news stories are on the same topic or in different topics.

In the first three steps of the experiments, we illustrated the realization task of SLD with independent decisions of these methods separately. However, in the next four steps (*4,5,6,7*) we made judgments using the OR-AND logical operators to obtain combined decision results for these methods. Thus, we had the chance to catch the relevant missed documents with

VSM by using the Named Entities methods [34]. So, in the fourth step, independent decisions about VSM word-based and NE intersection were carried out and coupling was done with OR logical operator. Following this, in the next step (fifth) we performed AND-logical combination of VSM (WB) and NE intersection methods. In the sixth and seventh steps, similar to two previous steps (*4,5*) experimental tests were carried out between entity-based VSM and NE intersection checking method.

In the final step, we used a resemblance function to calculate the similarity between two news stories using only named entity. Actually, by normalizing common entities between the stories, this function calculates the similarity score between news stories based on named entities. In order to reach the similarity score between news stories based on named entities, and also to emphasize the importance of the named entities in comparison, we used the resemblance function [5]. The resemblance function between two documents *a* and *b* is defined as follows:

$$f(a,b) = \frac{|a \bigcap b|}{|a \bigcup b|} \quad (5)$$

In Equation 5, the numerator is the number of entities common in *a* and *b* and the denominator is the number of all entities in *a* and *b*.

## IV. TESTING AND RESULTS

The dataset is divided into training (one third of the news stories) and test (two thirds of news stories) sets. The news stories in the training set are used to determine the threshold parameter value. In this respect, the threshold value of the VSM method was defined as the optimum point where recall and precision become equal. Tests were carried on the corpus which includes 3,922 news stories with known topic titles that were not used in the training set. During testing, news stories with known topic titles were compared with the rest of the news stories in the test set. Furthermore, logical OR/AND operators were applied to the results obtained using VSM and NE methods so that the effects of OR/AND operators on precision and recall measures could be evaluated. To determine the overall performance, precision, recall and F-measure values were also calculated. In the course of testing, Turkish stop-words were removed and also stemming was applied.

## V. DISCUSSION AND CONCLUSION

This paper analyzes how well the performance of the VSM is in the SLD task. Our purpose in this work was to detect whether two documents are linked or not. This paper presents a combination of different word-based and entity-based techniques to improve the performance of link detection. A combination of methods is shown to provide improved estimation performance in some cases.

The findings obtained using all methods are presented in Table I (*P: Precision, R: Recall, F: F-Measure, T: Threshold*). As the results indicate, the combinations using Boolean OR operator of VSM word-based and VSM entity-based with Named Entities intersection methods resulted in substantial improvements in performance.

The highest performance is obtained using the OR combination of VSM entity-based and NE intersection which was obtained with an F-measure value of 0.90 (recall: 0.84 and precision: 0.98). This combination achieved a substantial 30% increase in performance compared to the best case with VSM which resulted in an F-measure value of 0.60 (recall: 0.56 and precision: 0.65).

In this work we aimed at developing methods that provide higher precision and recall simultaneously. So, when we analyze the results of Named Entity Intersection (NEI) with a precision value of 0.13 and recall value of 0.81, we understand that this method was not very successful for the SLD task. But, these results can also be interpreted differently as it is possible to conclude that this method is able to determine that two articles are on different topics with a rate of 81%.

TABLE I. TEST RESULTS

| **Method** | **P** | **R** | **F** | **T** |
|---|---|---|---|---|
| VSM (WB) | 0.61 | 0.58 | 0.59 | 0.05 |
| VSM (EB) | 0.65 | 0.56 | 0.60 | 0.02 |
| NEI | 0.13 | **0.81** | 0.23 | - |
| VSM (WB) *OR* NEI | 0.77 | 0.75 | **0.76** | 0.04 |
| VSM (WB) *AND* NEI | 0.53 | 0.51 | 0.52 | 0.05 |
| VSM (EB) *OR* NEI | 0.98 | 0.84 | **0.90** | 0.01 |
| VSM (EB) *AND* NEI | 0.65 | 0.56 | 0.60 | 0.02 |
| NERF | 0.98 | 0.13 | 0.22 | 0.02 |

In this work, SLD that drew special attention within the TDT research is applied for the first time on a Turkish corpus using Named Entities method using named entities extracted from the text manually. The results clearly show that the VSM performance is substantially affected by the combination of NE methods, in identifying the similarities of news stories.

## VI. FUTURE WORK

Further studies should investigate the effect of Named Entities individually and in different combinations on the retrieval performance for identification. Actually, controlling named entities and identifying individual intersection in order to determine which named entities lead to better performances are two different approaches that can be investigated.

Furthermore, named entities with binary (person and location), triple (person, location, and date) or quad (person, location, date and organization) combination intersections between news can be analyzed. By analyzing these named entities, we were able to detect which combination of named entities is better for the SLD task. In some studies, event based methods are used for SLD detection. “Date” and “Location” named entities are used to extract events. Extraction of events will enable the use of event word-based methods.

 ISBN: 978-1-61208-457-2 

In this study, we do not use "query expansion" which is a well-established technique in IR. An important performance parameter in IR applications is the query length, i.e., the number of words used in queries. The effect of query length on retrieval performance based on named entities can also be analyzed for IR application environments. "Query expansion" technique based on non-entity words (all-words) was carried out by the same authors previously in [26] and [27].

## ACKNOWLEDGMENTS

This Work Is Partially Supported By The Scientific And Technical Research Council Of Turkey (TÜBİTAK) Under The Grant Number 111k030.

## REFERENCES

[1] Tonta, Y., Bitirim, Y., and Sever, H. (2002). Türkçe arama motorlarında performans değerlendirme. Total Bilişim.

[2] Allan, J. (2002). Introduction to topic detection and tracking. In *Topic detection and tracking* (pp. 1-16). Springer US.

[3] Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V. and Thomas, S. (2002, March). Relevance models for topic detection and tracking. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 115-121). Morgan Kaufmann Publishers Inc..

[4] Kose, G., Tonta, Y., Ahmadlouei, H., and Polatkan, A. C. (2013, November). Story Link Detection in Turkish Corpus. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on* (Vol. 1, pp. 154-158). IEEE.

[5] Allan, J., Carbonell, J. G., Doddington, G., Yamron, J. and Yang, Y. (1998). Topic detection and tracking pilot study final report.

[6] Letian Wang and Fang Li, Story Link Detection Based on Event Words, Springer-Verlag Berlin Heidelberg 2011

[7] Allan, J., Lavrenko, V. and Swan, R. (2002), Explorations Within Topic Tracking and Detection, Topic Detection and Tracking: Event-based Information Organization, J. Allan, Ed., Kluwer Academic Publishers, pp. 197-224.

[8] Chen, F., Farahat, A., Brants, T., Multiple similarity measures and sourcepair information in story link detection. Presented in the Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), Boston, Massachusetts, pp. 313–320, 2004.

[9] Allan, J., Lavernko, V., Nallapati, R. UMass at TDT2002. Presented in the Proceedings of the Topic Detection and Tracking Workshop, 2002.

[10] Allan, J., Lavrenko, V., Frey, D., Khandel Wal, V. UMass at TDT2000. Presented in the Proceedings of Topic Detection and Tracking Workshop, 2000.

[11] Chirag, S. and Koji E., (2009). Improving Document Representation for Story Link Detection by Modeling Term Topicality. IPSJ Online Transactions

[12] Shah, C., Croft, W. B. and Jensen, D. (2006). "Representing Documents with Named Entities for Story Link Detection (SLD)," a poster presentation at the ACM Fifteenth Conference on Information and Knowledge Management (CIKM) 2006, Arlington VA, November 6-11, 2006.

[13] Tadej Š. and Marko Grobelnik, (2009) Story Link Detection With Entity Resolution. ACM, Madrid, Spain

[14] Schultz, J. M. And Liberman, M. Y. (2002). Towards a "Universal Dictionary" for multi-language information retrieval applications. In *Topic detection and tracking* (pp. 225-241). Springer US.

[15] Makkonen, J., Ahonen-Myka, H. and Salmenkivi, M. (2003). Topic Detection and Tracking with Spatio-Temporal Evidence. In Proceedings of 25th European Conference on Information Retrieval Research (ECIR 2003). 251-265.

[16] Kumaran, G. and Allan, J. (2005). Using names and topics for new event detection. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 121-128). Association for Computational Linguistics.

[17] Makkonen, J., Ahonen-Myka, H. and Salmenkivi, M. (2002). Applying Semantic Classes in Event Detection and Tracking. Proc. International Conference on Natural Language Processing (ICON'02). 175-183.

[18] Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H. C. and Uyar, E. (2010). New event detection and topic tracking in Turkish. *Journal of the American Society for Information Science and Technology*, *61* (4), 802-819.

[19] Yang, Y., Carbonell, J., Brown, R., Lafferty, J., Pierce, T. Ault, T. (2002). Multi-strategy learning for topic detection and tracking. In J. Allan (Ed.), Topic Detection and Tracking: Event-based Information Organization (pp. 85-114). Norwell, MA: Kluwer Academic Publishers.

[20] Dalkiliç, F.E., Gelisli, S. and Diri, B. (2010). "Türkçe Kural Tabanli Varlik Ismi Tanima", (Turkish Rule Based Assets Recognition) 18. Sinyal Isleme ve Uygulama Kurultayi, Diyarbakir, (22-24 Nisan) 2010.

[21] Küçük, D. and Yazici, A. (2010). A Hybrid Named Entity Recognizer for Turkish with Applications to Different Text Genres. In Proceedings of the 25th International Symposium on Computer and Information Sciences (ISCIS). London, UK. E. Gelenbe et al. (Eds.): Computer and Information Sciences, LNEE 62, pp. 113-116.

[22] Xianshu Z. and Tim O., (2013), Finding News Story Chains Based on Multidimensional Event Profile.OAIR2013, Lisbon, Portugal

[23] Hua Zhao1 and Tiejun Zhao, (2009). Applying Dynamic Co-occurrence in Story Link Detection. Journal of Computing and Information Technology

[24] Rennie, J. D. M. (2008). Derivation of the F-measure, 2004. URL http://people. csail. mit. edu/jrennie/writing/fmeasure. pdf. accessed 15 May 2013.

[25] Nomoto, T. (2010). Two-tier similarity model for story link detection. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 789-798). ACM.

[26] Raghavan, V. V. and Sever, H. (1995). On the Reuse of Past Optimal Queries, Proceedings of 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, WA, USA, July 1995, pp. 344-351.

[27] Manmatha, R. and Sever, H (2002). A Formal Approach to Score Normalization for Metasearch, Human Language Technology Conference (HLT'02), March 24-27, 2002, San Diego, CA.

 ISBN: 978-1-61208-457-2 