Story Link Detection in Turkish Corpus

Güven Köse, Yaşar Tonta, Hamid Ahmadlouei
Department of Information Management
Hacettepe University
Ankara, Turkey
guvenkose@gmail.com
tonta@hacettepe.edu.tr
hamid2026@gmail.com

Aydın Can Polatkan
Center for Bioinformatics Tübingen
Faculty of Science, University of Tübingen
Tübingen, Germany
polatkan@informatik.uni-tuebingen.de

***Abstract*— Story Link Detection (SLD) is known as a sub-task of Topic Detection and Tracking (TDT). SLD aims to specify whether two randomly selected stories discuss the same topic or not. This sub-task drew special attention within the TDT research community as many tasks in TDT are thought to be solved automatically once SLD performs as expected. In this study, performance tests were carried out on the BilCol-2005 Turkish news corpus composed of approximately 209,000 news items using vector space model (VSM) and relevance model (RM) methods with respect to varied index term counts. Accordingly, best results obtained were as follows: the VSM method performed best with 30 terms (F-measure=0.2970) while RM method did with 4 terms (F-measure=0.1910). Furthermore, the combination of two methods using the AND and OR functions increased the precision ratio by 7.9% and recall ratio by 1.2%, respectively, indicating that retrieval performance of SLD algorithms can be increased to some extent by employing both VSM and RM models.**

Keywords*— *story link detection; topic detection and tracking; vector space model; relevance model.

I. INTRODUCTION

The exponential growth in the published online news items and the use of various channels, platforms and presentation styles make information seeking in the online age even more difficult than before [1][2]. More often than not, information tends to find users rather than users seeking information. Temporal Information Retrieval (T-IR) is one of the emerging areas of research in the field of IR. It aims to satisfy the temporal information needs of users by combining the traditional document relevance with time related relevance. Topic detection and tracking (TDT) is one of the sub-areas of T-IR. TDT algorithms based on IR models detect new unreported stories and organize them temporally, link incoming news items with previously detected stories on the same topic [3], and monitor news streams online till stories peter out [4]. Story link detection (SLD) algorithms play a key role in establishing linkages between stories discussing the same subjects [5][6]. Detected stories along with their updates can be delivered to the users based on their current interests and past behaviors. Story link detection, as defined as TDT sub-task, is the task of determining whether two stories, such as news, are about the same event, or linked.

In this paper, we analyze story link detection and investigate the effects of query expansion techniques. We present the performance of two different methods and show the improvements in the performance of the link detection system by the combination of these two methods. Based on our findings, results of combined methods for link generation is formulated, tested and presented.

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 that has been covered in the paper. In Section 4, we give details of the testing during our experiment and present their results. In Section 5, we present the conclusion by summarizing our contribution.

II. RELATED WORK

Information Retrieval Systems aim to find the information in documents in different environments in order to submit them to the interested users [7]. The basic functionality of an information retrieval system is required to meet the information needs of users, access to all relevant documents in the corpus and comb out non-relevant ones [8]. In order to determine relevant documents, an information retrieval system consists of a corpus of documents and a retrieval algorithm to compare the terms in the users' queries with the ones used to index the documents in the corpus. The concept of query expansion is a well-established technique in IR [26].

In recent years, academic studies in traditional information retrieval systems mainly focused on Topic Detection and Tracking (TDT) programs. The goal of the TDT is to develop technologies that organize, determine and follow-up the news stories of radio, newspapers, or television [6]. In order to accomplish this goal, a TDT algorithm analyzes the incoming news streams under five main tasks:

- Story Segmentation; to identify story boundaries automatically within a news stream.
- First Story Detection; to identify stories not encountered previously.
- Cluster Detection; to determine the subject(s) of the stories.
- Topic Tracking; to follow a story detected by the system.
- Story Link Detection; to distinguish if the two different stories are on the same subject or not.

In the TDT studies, the story link detection task is reported to have a critical role [5][6][9]. Carrying out the story link detection task successfully is expected to solve many problems in TDT [10]. Retrieval models used in story link detection are similar to that used in traditional IR systems (e.g., Boolean model [11], vector space model [12], probabilistic models [13][14][15], language model [16] and relevance model [17]). In TDT, story link detection task requires identifying pairs of linked stories for which he relevance model seems to work better than other models. The best current technology for link detection relies on the use of cosine similarity between document terms vectors with *tf.idf* term weighting. In a *tf.idf* model, the frequency of a term in a document (*tf*) is weighted by the inverse document frequency (*idf*), the inverse of the number of documents containing a term. Researchers have tested a number of similarity measures in the link detection task, including weighted sum, language modeling and Kullback-Leibler divergence, and found that the cosine similarity produced the best results [18]. In addition, using different methods together improved the retrieval performance [19][20][21][32][34][35].

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

In this context, this work investigates the story link detection performance of different retrieval functions and combinations thereof on a Turkish corpus. It concentrates on vector space and relevance models (and their combination) that have so far produced the optimum precision and recall values in the TDT studies.

III. METHODOLOGY

A. TDT Test Collection

In order to perform an experimental study, we used the new event detection and topic tracking test collection (BilCol-2005) developed by the Information Retrieval Group at Bilkent University [4]. The test collection contains 209,305 news items from five different Turkish news sources on the web, namely CNN Türk, Haber 7, TRT, and daily newspapers of Milliyet and Zaman. In the BilCol-2005 corpus, some 5,883 news items were classified under 80 different topic titles while the rest (203,442) have yet to be classified. In this work, tests were carried using the news items with the topic titles, assuming that the relevance of the rest of the news items to the classified ones were unknown.

B. Evaluation Methodology

We assessed the performance by computing the precision, recall and the F-measure that is based on precision and recall. Recall is the proportion of relevant documents retrieved and precision is the proportion of retrieved documents that are relevant [22]. F-measure identifies the harmonic mean of precision and recall [23]. In this work, we assume that high precision and high recall or high F-measure values represent better results.

C. SLD Methods Used In The Study

In story link detection task, many methods are used comparing the quantity of the overlapping words within the two stories. Large numbers of overlapping words between the two stories represent higher probability that the two stories discuss the same topic. This approach formed the basis of all the methods from vector space models [24][25][26][27] up to statistical language models [16][28][29]. Information retrieval researchers focus on 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. Depending on the studied areas, word-based methods [30] and language models [16] are usually used for the representation of the documents.

In this study, we used the vector space model (VSM) and relevance model (RM) to carry out the SLD task on the Turkish corpus. Although these methods have been widely used to solve the SLD problem in TDT studies, there is, to the best of our knowledge, no study carried out to test them on a Turkish corpus. [5][6][17][25][31].

The vector space model developed in the late 1960s is still a very popular approach and commonly used in IR systems as a retrieval function [17][24][25][30][32]. In this model, documents and queries are represented as vectors of index terms and similarity between these vectors prove the document/query matchup. Coefficients contained in the vectors highlight the importance of each term to what extent it represents the documents and/or queries.

In traditional IR methods, 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*) [33].

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

In TDT studies, the term vector is created for each document. Then, similarity between the two vectors (*a* and *b*) is calculated as in (1) where $tf_a(w)$represents the frequency of word *w* in *a* document, $tf_b(w)$ represents the frequency of word *w* in *b* document, and *idf(w)* represents the frequency of word *w* in all documents in the corpus.

Relevance model is the advanced version of Language Model that is used extensively in carrying out the story link detection task [3][5][17][31]. Relevance model offers a new approach to the estimation of probabilities when the necessary conditions of training data are absent. In a document related with a query, the probability of the word *w,* and *R* representing the set of relevant documents to the query is identified as the *P(w|R)* conditional probability. Accordingly, using *P(w|R)*, as the probability of the word in a collection, the maximum likelihood is predictable as in the following equation:

$$P(w \mid D) = \lambda P_{ml}(w \mid D) + (1-\lambda) P_{bg}(w) = \lambda \frac{tf_{w,D}}{|D|} + (1-\lambda) \frac{cf_w}{coll.size} \quad (2)$$

Using Equation 2, we created the topic model for each document. After this stage, the probability distributions of the two models were compared on the basis of Kullback-Leibler to determine document similarity [17].

IV. TESTING

The BilCol-2005 collection is divided into training (one third of the news items) and tests (two thirds of news items) sets. The news items in the training set obtained the threshold parameter value. In this respect, the threshold values for the VSM and RM methods were defined as the optimum point where recall and precision become equal. Tests were carried through the corpus with 3,922 news items with known topic titles and 135,609 news items with unknown topic titles that were not used as training documents. During testing, each news item with known topic titles was compared with the rest of the news items in the test set. For each query, a pairwise classification table is created.

In order to identify the effects of the number of index terms on the match performance, tests were repeated for 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 225, 250, 275, 300, 400, 500 and 1000 terms respectively. In addition, logical operators AND, and OR were applied on the results obtained through the use of VSM and RM methods so that the effects of Boolean operators AND and OR on precision and recall measures can be seen. In the last step, the micro-averaging method was used to create the common binary classification table of the entire test. To determine the overall performance, precision, recall and F-measure values were calculated. In the course of testing, Turkish stop-words were removed but stemming was not applied.

V. DISCUSSION AND CONCLUSION

Findings obtained using the vector space model and relevance model are shown in Figure 1 and Figure 2, respectively. In both figures horizontal axis represents the index terms. In vector space model, the best performance was obtained with 30 terms with an F-measure value of 0.2970 (recall: 0.2642, precision: 0.3393). In relevance model, the best performance was obtained with 4 terms with an F-measure value of 0.1910 (recall: 0.1625, precision: 0.2316). It appears that the selected best VSM method is more advantageous than the selected best RM method, providing higher recall (%10.17) and precision (%10.77) precision values.

Findings obtained with the combinations of Boolean operators AND and OR are shown in Figure 3 and Figure 4, respectively. The highest performance for AND combinations was obtained with 4 terms with an F-measure value of 0.2216 (recall: 0.1504 precision: 0.4183). The highest performance for OR combinations was obtained for 15 terms with an F-measure value of 0.2641 (recall: 0.2762 and precision: 0.2531). Accordingly, the AND combination of the methods achieved a %7.9 increase (VSM – 30 terms) compared to the best case with the highest precision value. Similarly, the OR combination of the methods achieved a %1.2 increase compared to the best case with the highest recall value.

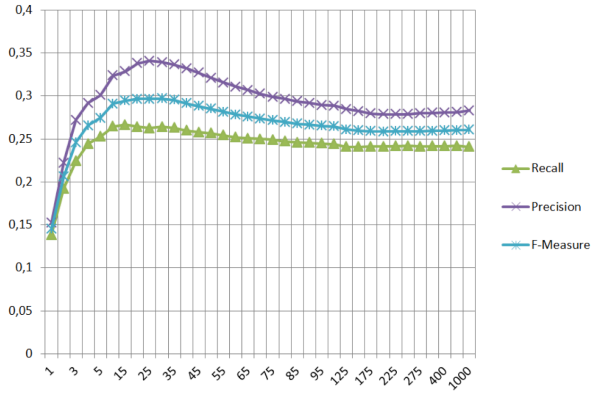

Fig. 1. Test results of VSM

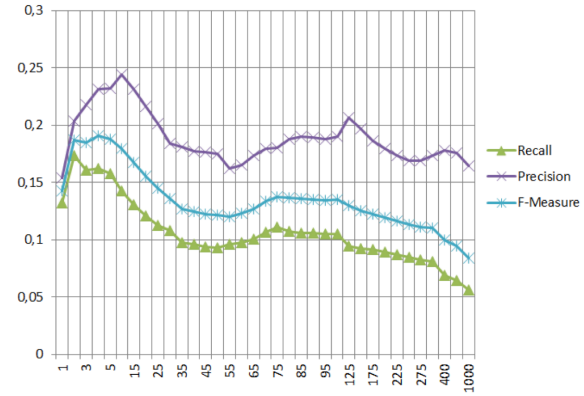

Fig. 2. Test results of RM

In this work, SLD that drew special attention within the TDT research is applied for the first time on a Turkish corpus using two different methods. Findings clearly show that VSM performed better than RM in identifying the similarities of news items. For further work, the effects of named entities on the retrieval performance for the identification of similar news will be studied and reported.

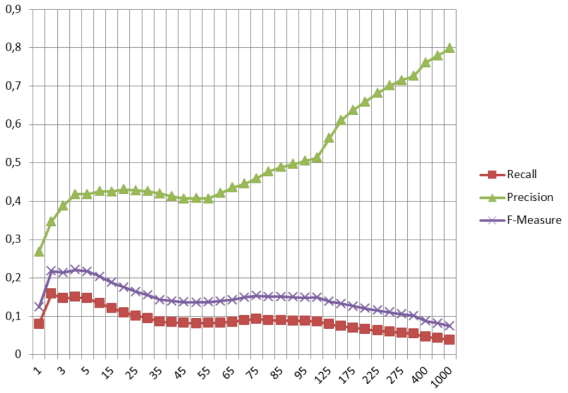

Fig. 3. AND combination of VSM and RM methods

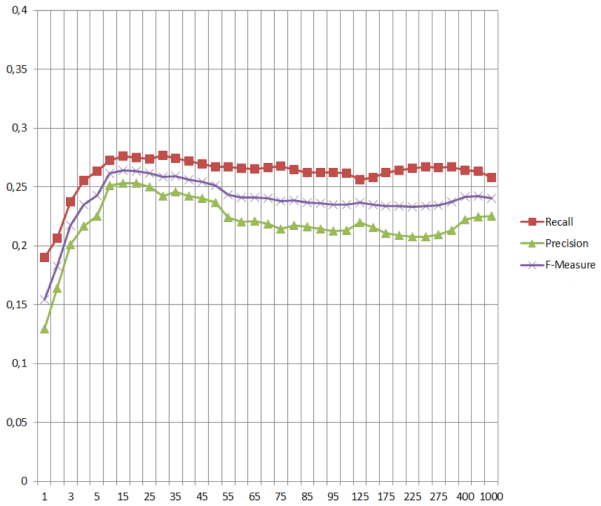

Fig. 4. OR combination of VSM and RM

ACKNOWLEDGMENT

This work is partially supported by the Scientific and Technical Research Council of Turkey (TÜBİTAK) under the grant number 106E014.

REFERENCES

[1] Lavrenko, V. (2009). *A generative theory of relevance* (Vol. 26). Springer.

[2] Yang, Y., Ault, T., Pierce, T., & Lattimer, C. W. (2000, July). Improving text categorization methods for event tracking. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 65-72). ACM.

[3] Chen, F., Farahat, A., & Brants, T. (2004, May). Multiple similarity measures and source-pair information in story link detection. In *proceedings of HLT-NAACL* (pp. 313-320).

[4] Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H. C., & 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.

[5] Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., & 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..

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

[7] Meadow, C. T., Kraft, D. H., & Boyce, B. R. (1999). *Text information retrieval systems*. Academic Press, Inc..

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

[9] Allan, J., Papka, R., & Lavrenko, V. (1998, August). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 37-45). ACM.

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

[11] Robertson, S. E. (1977). Theories and models in information retrieval. *Journal of Documentation*, *33*(2), 126-148.

[12] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

[13] Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, *7*(3), 216-244.

[14] Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, *35*(3), 243-255.

[15] Sparck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, *36*(6), 779-808.

[16] Ponte, J. M., & Croft, W. B. (1998, August). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 275-281). ACM.

[17] Lavrenko, V., & Croft, W. B. (2001, September). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 120-127). ACM.

[18] Allan, J., Lavrenko, V., & Jin, H. (2000, November). First story detection in TDT is hard. In *Proceedings of the ninth international conference on Information and knowledge management* (pp. 374-381). ACM.

[19] Hatzivassiloglou, V., Gravano, L., & Maganti, A. (2000, July). An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 224-231). ACM.

[20] Kumaran, G., & Allan, J. (2004, July). Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 297-304). ACM.

[21] Kumaran, G., & Allan, J. (2005, October). 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.

[22] Van Rijsbergen, C. J. (1979). *Information Retrieval*. Dept. of Computer Science, University of Glasgow. *URL: citeseer. ist. psu. edu/vanrijsbergen79information. html*. accessed 15 May 2013.

[23] 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.

[24] Frakes, W. & Baeza Yates, R. (1992). *Information Retrieval: Data Structure and Algorithm, Clustering Algorithms*, Prentice-Hall, Englewood Cliffs.

[25] Schultz, J. M., & Liberman, M. (1999, February). Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of the DARPA broadcast news workshop* (pp. 189-192). San Francisco: Morgan Kaufmann.

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

[27] Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, *18*(1), 79-112.

[28] Berger, A., & Lafferty, J. (1999, August). Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 222-229). ACM.

[29] Miller, D. R., Leek, T., & Schwartz, R. M. (1999, August). A hidden Markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 214-221). ACM.

[30] Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley. Reading, Mass.

[31] Leek, T., Schwartz, R., & Sista, S. (2002). Probabilistic approaches to topic detection and tracking. In *Topic detection and tracking* (pp. 67-83). Springer US.

[32] Nomoto, T. (2010, October). 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.

[33] Salton, G. & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill Book Co., New York.

[34] Chen, F., Farahat, A., & Brants, T. (2003, May). Story link detection and new event detection are asymmetric. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003*--short papers-Volume 2 (pp. 13-15). Association for Computational Linguistics.

[35] Farahat, A., Chen, F., & Brants, T. (2003, July). Optimizing story link detection is not equivalent to optimizing new event detection. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*-Volume 1 (pp. 232-239). Association for Computational Linguistics.