NAMED ENTITY RECOGNITION IN TURKISH USING ASSOCIATION MEASURES

Senem Kumova Metin[1], Tarık Kışla[2] and Bahar Karaoğlan[3]

[1]Department of Software Engineering, Izmir University of Economics, Izmir, Turkey
senem.kumova@ieu.edu.tr

[2]Department of Computer Education and Instructional Technologies, Ege University, Izmir, Turkey
tarik.kisla@ege.edu.tr

[3]International Computer Institute, Ege University, Izmir, Turkey
bahar.karaoglan@ege.edu.tr

ABSTRACT

Named Entity Recognition which is an important subject of Natural Language Processing is a key technology of information extraction, information retrieval, question answering and other text processing applications. In this study, we evaluate previously well-established association measures as an initial attempt to extract two-worded named entities in a Turkish corpus. Furthermore we propose a new association measure, and compare it with the other methods. The evaluation of these methods is performed by precision and recall measures.

KEYWORDS

Turkish, Point-wise Mutual Information, Mutual Dependency, First Kulczynski, Joint Probability

1. INTRODUCTION

Named Entity Recognition (NER) aims to locate and classify named entities in an unstructured text. NER classify elements in text into predefined categories such as the names of persons, organizations, locations etc. NER systems serves as an important pre-processing system for tasks such as information extraction, information retrieval, question answering and other text processing applications.

In NER systems, linguistic grammar-based techniques and/or statistical models have been used. Although grammar-based systems typically obtain better precision, these systems have lower recall and generally they are not independent of language. On the other hand, statistical NER systems require a large amount of manually annotated training data.

There has been a lot of work on NER systems on several languages especially on English. Although new models and techniques have been suggested, developing NER systems for Turkish is still a difficult process because of the agglutinative structure of the language.

In this study, we evaluate previously well-established association measures in order to extract two-worded named entities in Turkish corpus. The association measures are simply used to evaluate the relations between two or more items in a given set. We assume that the words in

DOI : 10.5121/acij.2012.3406

multi-worded named entities have such strong ties that these words tend to co-occur frequently. The methods/measures implemented in the study are based on point-wise mutual information [1], log frequency biased mutual dependency [2], first Kulczynski [3] and joint probability. Additionally, a new association measure is proposed and compared with other techniques.

The rest of this paper is organized as follows: Section 2 summarizes previous related work. Section 3 describes association measures, presents base set, and explains the experiments. Section 4 reports results, and concludes with further future work.

2. Related Work

In the 1990s, the name entity recognition was firstly described in the Message Understanding Conference (MUC), which was sponsored by the US Defense Advanced Research Projects Agency (DARPA). One of the first research papers in the field was presented by Lisa F. Rau [4]. Since then, many methods and strategies for automatic identification of named entities have been proposed. Methods for NER systems classify into 3 groups: the rule based approach, probabilistic approach, and the hybrid approach.

In the rule-based approach, the natural language descriptions and rules need to be formulated. The rules are used to define name entities using their syntactic and lexical structure with the help of manually annotated corpora. In addition to rules, rule-based approaches require gazetteers and general dictionary [5]. Differently from rule-based approach, the probabilistic approach does not require any natural language information. This approach builds their models by learning patterns from the annotated corpora [5]; and the approach displays good enough performance with large corpora ([6][7]). In [8] it is indicated that recent studies about NER are mostly based on probabilistic methods.

Although some studies address language independence or multilingualism in NER solutions, a large part of the NER studies are on English. NER studies on other languages than English have been also carried out; such as German ([9]) Spanish ([10]), Japanese ([11][12][13]), Chinese ([14] [15][16])), French ([17][18]), Greek ([19]), Italian ([20][21]), Bulgarian ([22]), Hindi ([23]), Polish ([24]), Russian ([25]), Swedish ([26]), Portuguese ([27]).

There is a limited work on NER systems applied on Turkish texts. To our best knowledge, the study of Cucerzan and Yarowski ([22]) is the first study on Turkish NER. In [22], a language independent bootstrapping algorithm that learns from word internal and contextual information of entities is presented and the proposed algorithm is experimented on five languages including Turkish. Following this, in [28] a statistical approach (HMMs) for NER is used on Turkish texts. In [29], a huge database of person, organization, and location names is constructed instead of employing a complex name entity extraction scheme. As a recent study, Kucuk and Yazici [30] presented a rule-based NER system for Turkish.

3. Experiments

3.1. Methods: Association Measures

In this section, detailed information about the proposed method and other competing methods; point-wise mutual information, log frequency biased mutual dependency, First Kulczynski and joint probability; are given.

Point-wise mutual information (*PMI*) is the association measure which generates a score depending on the mutual dependence of the two or more words. For two words x and y occurring consecutively in the corpus the *PMI* may be given as

$$PMI(xy) = \log_2 \frac{P(xy)}{P(x)*P(y)} \quad (1)$$

P(x), *P(y)* are the probabilities of the words *x* and *y* appearing separately respectively. *P(xy)* is the probability of *x* and *y* coming together.

Log frequency biased mutual dependency (*MD*) is a derived measure of mutual information ([31]). *MD* may be formulized for the words *x* and *y* as

$$MD(xy) = \log \frac{P(xy)^2}{P(x)*P(y)} + \log P(xy) \quad (2)$$

First Kulczynski (*FK*) gives a measure for association between two consecutive words in the corpus. *FK* for the words *x* and *y* is presented as

$$FK(xy) = \frac{f(xy)}{f(x\bar{y}) + f(\bar{x}y)} \quad (3)$$

where *f(xy)* is the co-occurrence frequency of the words x and y. $f(x\,\bar{y})$is the frequency of *x* followed by any word other than *y*. $f(\bar{x}y)$ is the occurrence frequency of *y* not preceded by *x*.

Joint probability (*JP*) is the probability of the words x and y to occur together in the corpus (*P(xy)*) and the method is accepted to be the easiest way to score the associations between words in a text.

The idea behind the proposed method (*PM*) in the study may be given as follows

> *"If the words x and y compose a two-word named-entity, the number of separate occurrences of x and y (f(x) and f(y) respectively) must be close or equal to the number of co-occurrences of the words; f(xy); together."*

The proposed method measures the distance of separate frequencies of words (*f(x)* and *f(y)*) from the co-occurrence frequency (*f(xy)*). The distances for the first and the second words are squared and summed to generate a single score for each named-entity candidate. The method is formulized below.

$$PM(xy) = \left(\frac{f(xy)-1}{f(x)-1}\right)^2 + \left(\frac{f(xy)-1}{f(y)-1}\right)^2 \quad (4)$$

3.2. Base Set

The experiment is utilized on Bilkent corpus which is compiled in Bilkent University for linguistics studies [28]. In the study, a base set is generated from the corpus as in [32]. The base set is constructed by merging the best 200 candidates from the ranked lists of competing association measures. The base set is re-ranked by each method and the ranked lists are used to evaluate the performance of each method.

3.3. Evaluation

The evaluation of competing methods is performed by precision and recall measures. The precision is simply the proportion of true named-entities to the total number of candidates in the set. The recall is the proportion of true named-entities to the total number of true named-entities in the base set. Instead of giving a single score of precision and recall for the whole set or any proportion of the set, we plot curves of precision and recall in order to present scores for any proportions of the whole set.

4. RESULTS AND CONCLUSION

The implementation of association measures on the Bilkent corpus has returned a base set of 569 two-worded candidates. The base set involves 32.34% true named-entities of two words. Figure 1 gives the precision curves for the competing methods: *PMI*, *MD*, *FK* and *JP*.

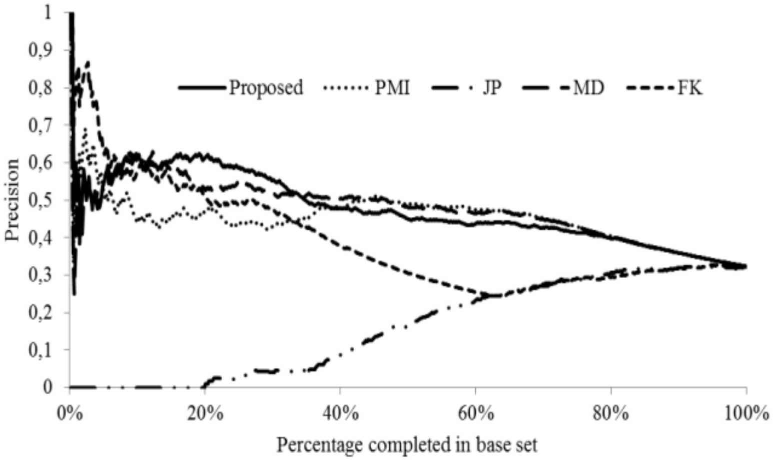

Figure 1. Precision curves of association measures in named entity recognition

It is observed from Figure 1 that any of the methods including the proposed method does not perform consistently better than the other. It is also seen that in the first ~10% percent of the base set, all the methods generate non-consistent precision values; between 20% and %40 proportions of the set the proposed method generates higher scores. In the range 40%-80%, the curves of mutual dependency and point-wise mutual information dominate the others. It is noteworthy that joint probability is the method which gives worst results in named-entity recognition.

Figure 2 depicts recall curves for the methods. Recall curves show that the competing methods except joint probability and first Kulczynski generate similar results of recall meaning that none of the methods retrieve true entities faster than others.

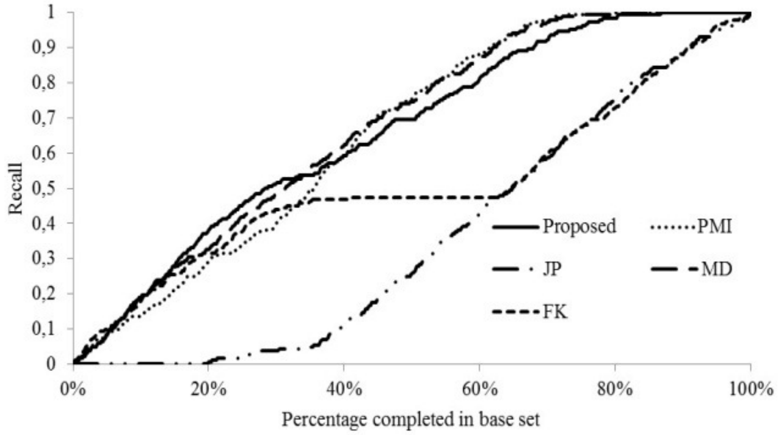

Figure 2. Recall curves of association measures in named entity recognition

5. CONCLUSIONS

As a conclusion, the attempt to evaluate several association measures including the proposed measure on named entity recognition showed that current association measures do not succeed in the complete base set. Since point-wise mutual information, mutual dependency and the proposed method achieve in several proportions of base set, we believe that if these methods are modified, some improvement in the results may be achieved.

REFERENCES

[1] K. Church, P. Hanks, (1990). Word association norms, mutual information and lexicography. Computational Linguistics, 16(1), 22–29.

[2] B. Daille, (1994). Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et _ltres linguistiques. PhD thesis, Universite Paris

[3] S. Kulczynski, (1927). "Classe des Sciences Mathématiques et Naturelles" Bulletin International de l'Acadamie Polonaise des Sciences et des Lettres Série B (Sciences Naturelles)(Supplement II): 57-203.

[4] L. F. Rau, 1991. Extracting Company Names from Text. In Proc. Conference on Artificial Intelligence Applications of IEEE.

[5] H. N. Traboulsi, "Named Entity Recognition: A Local Grammarbased Approach," unpublished Ph.D. dissertation, Department of Computing School of Electronics and Physical Sciences, University of Surrey Guildford, Surrey GU2 7XH, U.K, 2006

[6] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a High-Performance Learning Name-finder," in Proceedings of the fifth conference on Applied natural language processing, pp. 194 - 201, 1997.

[7] Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "NYU: Description of the MENE Named Entity System as Used in MUC-7," in Message Understanding Conference (MUC-7), 1998.

[8] D. Nadeau & S. Sekine, "A survey of named entity recognition and classification,“ Ed. S. Sekine and E. Ranchhod, unpublished Technical Report, 2007.

[9] C. Thielen,. 1995. An Approach to Proper Name Tagging for German. In Proc. Conference of European Chapter of the Association for Computational Linguistics.

[10] X. Carreras; L. Márques, L. Padró. 2003. Named Entity Recognition for Catalan Using Spanish Resources. In Proc. Conference of the European Chapter of Association for Computational Linguistic.

[11] M. Asahara, Y. Matsumoto, 2003. Japanese Named Entity Extraction with Redundant Morphological Analysis. In Proc. Human Language Technology conference – North American chapter of the Association for Computational Linguistics.

[12] S. Sekine, 1998. Nyu: Description of the Japanese NE System Used For Met-2. In Proc. Message Understanding Conference.

[13] S. Sekine, Isahara, H. 2000. IREX: IR and IE Evaluation project in Japanese. In Proc. Conference on Language Resources and Evaluation.

[14] L-J Wang, W-C Li, C-H, Chang. 1992. Recognizing Unregistered Names for Mandarin Word Identification. In Proc. International Conference on Computational Linguistics.

[15] H. H. Chen, J. C. Lee, 1996. Identification and Classification of Proper Nouns in Chinese Texts. In Proc. International Conference on Computational Linguistics.

[16] S. Yu, S. Bai, P. Wu. 1998. Description of the Kent Ridge Digital Labs System Used for MUC-7. In Proc. Message Understanding Conference.

[17] G. Petasis, F. Vichot, F. Wolinski, G.Paliouras, V. Karkaletsis, C.D. Spyropoulos. 2001. Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. In Proc. Conference of Association for Computational Linguistics.

[18] T. Poibeau. 2003. The Multilingual Named Entity Recognition Framework. In Proc.Conference on European chapter of the Association for Computational Linguistics.

[19] S. Boutsis; I. Demiros, V. Giouli, M. Liakata, H. Papageorgiou, S. Piperidis. 2000. A System for Recognition of Named Entities in Greek. In Proc. International Conference on Natural Language Processing.

[20] W.J. Black, F. Rinaldi, D. Mowatt. 1998. Facile: Description of the NE System used for Muc-7. In Proc. Message Understanding Conference.

[21] A. Cucchiarelli, P. Velardi, P. 2001. Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence. Computational Linguistics 27:1.123-131, Cambridge: MIT Press.

[22] S. Cucerzan, D. Yarowsky. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In Proc. Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[23] J. May, A. Brunstein, P. Natarajan, R.M. Weischedel. 2003. Surprise! What's in a Cebuano or Hindi Name? ACM Transactions on Asian Language Information Processing 2:3.169-180, New York: ACM Press.

[24] J. Piskorski. 2004. Extraction of Polish Named-Entities. In Proc. Conference on Language Resources an Evaluation.

[25] B. Popov, A. Kirilov, D. Maynard, D. Manov. 2004. Creation of reusable components and language resources for Named Entity Recognition in Russian. In Proc. Conference on Language Resources and Evaluation.

[26] D. Kokkinakis. 1998., AVENTINUS, GATE and Swedish Lingware. In Proc. of Nordic Computational Linguistics Conference.

[27] D. D. Palmer, S. Day. 1997. A Statistical Profile of the Named Entity Task. In Proc. ACL Conference for Applied Natural Language Processing.

[28] G. Tür, D. Hakkani-Tür, and K. Oflazer. 2003. A Statistical Information Extraction System for Turkish. Natural Language Engineering. Vol. 9 No 2 181-210

[29] K. Oflazer, Ö. Çetinoğlu, B. Say, “Integrating Morphology with Multiword Expression Processing in Turkish,” in Second ACL Workshop on Multiword Expressions: Integrating Processing, pp. 64 - 71, 2004.

[30] D. Kucuk, A. Yazici. 2009. Named entity recognition experiments on Turkish texts. In Proceedings of the 8th International Conference on Flexible Query Answering Systems, FQAS ’09, pages 524–535, Berlin, Heidelberg. Springer-Verlag.

[31] C. Hore, M. Asahara,, Y. Matsumoto. 2005. Automatic Extraction of Fixed Multiword Expressions. IJCNLP 2005: 565-575

[32] S. Kumova Metin., B. Karaoğlan, 2011. Measuring Collocation Tendency of Words, Journal of Quantitative Linguistics, Vol 18/2

Authors

Senem Kumova Metin has taken her B.S degree from Electrical and Electronics Engineering Department, Ege University (2001), M.S. and PhD degrees from International Computer Institute, Ege University, İzmir (2005, 2011). She is mainly interested in Natural Language Processing applications. She is currently working as an Assistant Professor in İzmir University of Economics. She gives courses on programming languages and data structures.

Tarık Kışla has taken his B.S. degree from Mathematics, Ege University (1998), and M.S and Phd. degrees from International Computer Institute. He is currently working as a LECTURER in Department of Computer and Instructional Technologies, Ege University. He gives lectures on information technologies, algorithms, programming language, databases, web design and networks. Further information about his publications and projects can be found from http://egitim.ege.edu.tr/~tkisla

Bahar Karaoğlan has taken her B.S. degree from Electrical and Electronics Engineering, Bogazici University (1977), and M.S. degree from Computer Science, Bogaziçi University (1979), and Phd. Degree from Computer Engineering, Ege University (1991). She is a PROFESSOR and Vice Director of International Computer Institute of Ege University, İzmir, Turkey; Turkish Scientific Committee Head in EU MedNet’U (Mediterenean Network of Universities) project; project leader in several national projects funded by Scientific and Technological Research Council of Turkey (TÜBİTAK), and Ege University Science and Technology Application and Research Center. She is giving information retrieval, multimedia systems, computer architecture, information systems and expert systems courses. Further information about her publications and projects can be found from http://ube.ege.edu.tr/~bahar