TEXT SEGMENTATION USING NAMED ENTITY RECOGNITION AND CO-REFERENCE RESOLUTION

Pavlina Fragkou
Technological Educational Institution of Athens (TEI-A), Dept. of Library Science and Information Systems
Ag. Spyridonos, 122 10 Egaleo, Greece

Keywords: Text segmentation, Named entity recognition, Co-reference resolution, Information extraction.

Abstract: In this paper we examine the benefit of performing named entity recognition and co-reference resolution to a benchmark used for text segmentation. The aim here is to examine whether the incorporation of such information enhances the performance of text segmentation algorithms. The evaluation using three well known text segmentation algorithms leads to the conclusion that, the benefit highly depends on the segment's topic, the number of named entity instances appearing in it, as well as the segment's length.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

The information explosion of the web aggravates the problem of effective information retrieval. To address this, various techniques such as text segmentation and information extraction provide partial solutions to the problem. More specifically, text segmentation methods are useful in identifying the different topics that appear in a document. On the other hand, information extraction methods try to identify portions of text that refer to a specific topic, by focusing on the appearance of instances of specific types of named entities (such as person, date, location, etc.) according to the thematic area of interest.

The question that arises is whether the combination of text segmentation and information extraction (and most specifically the named entity recognition and co-reference resolution steps) can prove to be beneficial for the identification of the various topics that appear in a document.

This paper examines the benefit of performing named entity recognition and co-reference resolution in the Choi's corpus (Choi, 2000). This corpus is used by researchers as benhmark for examining the performance of text segmentation algorithms. It must be stressed that, the focus is not on finding the algorithm that achieves the best segmentation performance on the corpus, but on the benefit of performing named entity recognition as well as co-reference resolution on a corpus used for text segmentation.

The structure of the paper is as follows. Section 2 provides an overview of related methods. Section 3 presents the steps performed for the creation of the "annotated" corpus. Section 4 presents evaluation results obtained by using three well known text segmentation algorithms, while Section 5 provides conclusions and future steps.

2 RELATED WORK

The text segmentation problem of concatenated text can be stated as follows: given a text which consists of several parts (each part dealing with a different subject), it is required to find the boundaries between the parts. A starting point to this is the calculation of the within-segment similarity based on the assumption that, parts of a text having similar vocabulary are likely to belong to a coherent topic segment. It must be stressed that, within-segment similarity is calculated on the basis of words but not on the basis of the application of other more sophisticated techniques such as named entity recognition or co-reference resolution. In the literature, several word co-occurrence statistics are proposed (Choi, 2000), (Choi et al., 2001), (Hearst, 1997), (Utiyama and Isahara, 2001). A significant difference between text segmentation methods is that, some authors evaluate the similarity between *all* parts of a text (Choi, 2000), (Choi et al., 2001), (Ponte and Croft, 1997), (Reynar, 1994), (Xiang and Hongyuan, 2003), while other between adjacent

Fragkou P..
TEXT SEGMENTATION USING NAMED ENTITY RECOGNITION AND CO-REFERENCE RESOLUTION.
DOI: 10.5220/0003181603490354
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 349-354
ISBN: 978-989-8425-40-9
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

parts (Hearst, 1997), (Heinonen, 1998). To penalize deviations from the expected segment length, several methods use the notion of "length model" (Heinonen, 1998), (Ponte and Croft, 1997). Dynamic programming is often used in order to calculate the globally minimal segmentation cost (Heinonen, 1998), (Reynar, 1994), (Xiang and Hongyuan, 2003), (Kehagias et al., 2004), (Qi et al., 2008). Current approaches involve the improvement of the dotplotting technique (Yen et al., 2005), the improvement of Latent Semantic Analysis (Bestgen, 2006) and the improvement of Hearst's TextTiling method (Hearst, 1997) presented by (Kern and Granitzer, 2009).

Information extraction, from a different point of view, aims to locate within a text passage domain-specific and pre-specified facts (e.g., in a passage about athletics, facts about the athlete participating in a 100m event, such as name, nationality, performance, as well as facts about the specific event, like the event name). More specifically, information extraction is about -among others- extracting from texts: (a) *Entities*: textual fragments of particular interest, such as persons, places, organizations, dates, etc. (b) *Mentions*: the identification of all lexicalisations of an entity in texts. For example, the name of a particular person can be mentioned in different ways inside a single document, such as "Lebedeva", "Tatiana Lebedeva", or "T. Lebedeva". The following pre-processing steps are applied in order to perform information extraction: (a) *Named Entity Recognition,* where entity mentions are recognized and classified into proper types for the thematic domain in question (b) *Co-reference*, where all the mentions that represent the same entity are identified and grouped together according to the entity they refer to.

Co-reference resolution complementary includes the step of anaphora resolution. The term anaphora denotes the phenomenon of referring to an entity already mentioned in a text -most often with the help of a pronoun or a different name. Co-reference basically involves the following steps: (a) pronominal co-reference (which is about finding the proper antecedent for personal pronouns), possessive adjectives, possessive pronouns, reflexive pronouns and pronouns this and that (b) identification of cases where both the anaphor and the antecedent refer to identical sets or types. This identification requires some world knowledge or specific domain knowledge. It also includes cases such as reference to synonyms or the case where the anaphor matches exactly or is a substring of the antecedent (c) ordinal anaphora for cardinal numbers and adjectives such as "former" and "latter".

The importance of text segmentation and information extraction is apparent in a number of applications, such as noun phrase chunking, tutorial dialogue segmentation, focused crawling, text summarization, semantic segmentation and web content mining. In (Fragkou, 2009) the potential use of text segmentation in the information extraction process was examined. In this paper the reverse problem is examined i.e., the use of information extraction techniques in the text segmentation process. Those techniques are applied on a benchmark used for text segmentation, resulting in the creation of an "annotated" corpus. Evaluation was performed using three well-known segmentation algorithms (Choi et al., 2001), (Kehagias et al., 2004) and (Utiyama and Isahara, 2001) applied both in the original as well as the "annotated" corpus.

A similar work was presented in (Sitbon and Bellot, 2005). The authors used two corpora. The first one was a manually-built, French-news corpus which contained four series of 100 documents, where each document was composed of ten segments extracted from "Le Monde" journal. The second one was referring to a single topic (sport). In each of those corpora, they performed named entity recognition using three types of named entities: person name, location, and organization. The authors state use of anaphors but provide no further details. They used named entity instances as components of lexical chains to perform text segmentation. Their results showed that, the use of named entities does not improve segmentation accuracy.

3 METHOD

Existing algorithms performing text segmentation exploit a variety of word co-occurrence statistic techniques in order to calculate the homogeneity between segments, where each segment refers to a single topic. However, they do not exploit the importance that several words may have in a specific context. Examples of such words are person names, locations, dates, group of names, scientific terms etc. The importance of those terms is further diminished by the application of word processing techniques, i.e., stop list removal and stemming on words such as pronouns or adjectives. We aim to exploit whether the identification of such words can be beneficial for the segmentation task. This identification requires the application of named entity recognition and co-reference resolution thus,

their (manual or not) annotation effort is under examination.

Our work differs from the one presented in (Sitbon and Bellot, 2005) in the following points: (a) we use a widely accepted benchmark i.e., Choi's text segmentation corpus (Choi, 2000) (b) we use an additional named entity i.e., date (c) we perform manually co-reference resolution (i.e., all the aforementioned tasks of co-reference resolution) complementary to named entity recognition to those portions of text that refer to named entity instances (d) the produced "annotated" corpus was evaluated using three text segmentation algorithms.

3.1 The Corpus

The corpus used here is the one generated by Choi (Choi, 2000). The description of Choi's 700 samples corpus is as follows: "*A sample is a concatenation of ten text segments. A segment is the first n sentences of a randomly selected document from the Brown Corpus. A sample is characterized by the range n*." Table 1 gives the corpus statistics per dataset.

Table 1: Test Corpus Statistics per dataset (Choi, 2000).

Range of n	3-11	3-5	6-8	9-11
#samples	400	100	100	100

More specifically, Choi created his corpus by using sentences selected from 44 documents belonging to category A *Press* and 80 documents belonging to category J *Learned*. The description of Brown Corpus states that category A contains documents about *Political, Sports, Society, Spot News, Financial and Cultural*. Category J contains documents about *Natural Sciences, Medicine, Mathematics, Social and Behavioral Sciences, Political Science, Law, Education, Humanities, Technology and Engineering*. Documents belonging to category J usually contain portions of scientific publications about mathematics or chemistry. Thus, they contain scientific terms such as *urethane foam, styrenes, gyro-stabilized platform system* etc. On the other hand, the majority of documents of category A usually contain person names, locations, dates, groups of names etc.

3.2 Named Entity Annotation

A number of annotation tools exist in the literature such as GATE (http://gate.ac.uk/), Callisto (http://callisto.mitre.org/), MMAX2 (Müller and Strube, 2006), AeroSWARM (Corcho, 2006), Knowtator (Ogren, 2006), Ellogon (Petasis, 2003), and Wordfreak (http://wordfreak.sourceforge.net/). However the majority of those tools require training, which is usually focused on a single topic. The important number of different topics appearing in the 124 documents of the Brown Corpus precludes the creation of training models (one for each topic) leading us to perform manual annotation. Thus, we performed manual named entity recognition and co-reference resolution on each of the 10 segments of the 700 samples. In order to cover the majority of entities and mentions in each segment, we selected four types of named entities: person name, location, date, and group name. The most general type is that of group name, which is used for the annotation of words and terms that do not fall into the other categories. It was also used for the annotation of scientific terms frequently appearing in segments.

We note that in Semcor (http://multisemcor.itc.it/semcor.php) a different annotation for the majority of documents of category A and J was performed. Most specifically, *"The Semcor corpus is composed of 352 texts. In 186 texts, all open class words (nouns, adjectives and adverbs) are annotated with PoS, lemma and sense according to Princenton Wordnet 1.6, while in the remaining 166 text only verbs are annotated with lemma and sense*". This type of annotation differs from the one performed here. More specifically, even though in Semcor nouns are classified into three categories (person name, group, and location), identification of identical named entity instances as well as mentions resulting from the application of co-reference resolution is not performed. Additionally, Semcor does not provide annotations for all documents belonging to category J nor for all named entity instances (as for example scientific terms like *urethane foam*).

Consequently, in each segment manual named entity annotation of proper names belonging to one of the four categories was performed. The annotation took under consideration the assignment of lemmas to categories for the cases of person name, group and location appearing in Semcor. We believe that the substitution of words with named entity instances does not have an effect in the performance of a segmentation algorithm. Based on this, during manual named entity annotation, we additionally: (a) substituted every reference of the same instance with the same named entity identifier. For example in the sentences "*James P. Mitchell and Sen. Walter H. Jones R-Bergen, last night disagreed on the value of using as a campaign issue a remark by Richard J. Hughes,... . Mitchell was for using it, Jones against"*, we first identified three instances of

person names. We further used the same entity identifier for *James P. Mitchell* and *Mitchell* and the same entity identifier for *Sen. Walter H.Jones R-Bergen* and *Jones* (b) we substituted every reference of the same instance, resulted from co-reference resolution, with the same named entity identifier (for example in the sentences "*Mr. Hawksley, the state's general treasurer,... He is not interested in being named a full-time director*", we substituted *He* with the named entity identifier given to *Mr. Hawksley*).

In align with Secmor, group names involved expressions such as "*House Committee on Revenue and Taxation*" or "*City Executive Committee*". The annotation of location instances included possible derivations of them such as "*Russian*". The annotation of date instances included both simple date form (consisting only of the year or month) and more complex forms (containing both month, date and year). It must be stressed that, co-reference resolution was performed only on portions of text that refer to named entity instances and not on the text as a whole. This assumption makes manual annotation more attractable than the use of co-reference resolution tools like Link Grammar Parser (http://www.link.cs.cmu.edu/link/) or YamCha (http://chasen.org/~taku/software/yamcha/).

The annotation process led to the conclusion that, segments belonging to category A contain on average, more named entity instances compared to those belonging to category J. The difference in the results is highly related to the topic discussed in every segment of each category. More specifically, the largest part used as segment (i.e., portions of 11 sentences) in the Choi's benchmark, from each of the 124 documents of the Brown corpus was selected. After that, the minimum, maximum, and average number of named entity instances appearing in them, were calculated. The results are listed in Table 2.

Table 2: Statistics regarding the number of named entity instances appearing in segments of Category A and J.

Category/ NE instances per segment	Min	Max	Average
Segments of Category A	2	53	28.318
Segments of Category J	2	57	18.400

4 EVALUATION

The "annotated" corpus that resulted from the previously described process was evaluated using three text segmentation algorithms. The first is Choi's C99b (Choi, 2001), which creates a similarity matrix for sentences appearing in a text using Latent Semantic Analysis. C99b then finds topic boundaries by recursively seeking the optimum density along the matrix diagonal. The second algorithm is the one proposed by (Utiyama and Isahara, 2001). This algorithm finds the optimal segmentation of a given text by defining a statistical model which calculates the probability of words to belong to a segment. To find the maximum probability segmentation, it calculates the minimum-cost segmentation obtained by the minimum cost path in a graph. Both algorithms benefit from the fact that, they do not require training and they are publicly available.

The third algorithm used is introduced by (Kehagias et al., 2004) which, contrary to the previous ones, requires training. More specifically, this algorithm uses dynamic programming to find both the number and the location of segment boundaries. The algorithm decides the locations of boundaries by calculating the globally optimal splitting (i.e., global minimum of a segmentation cost) on the basis of a similarity matrix, a preferred fragment length, and a defined cost function.

4.1 Experiments - Results

We evaluate the performance of the algorithms in the original and "annotated" corpus using three widely known indices: Precision, Recall and Beeferman's *Pk* metric (Beeferman at al., 1999). Precision is defined as *"the number of the estimated segment boundaries which are actual segment boundaries" divided by "the number of the estimated segment boundaries*". Recall is defined as "*the number of the estimated segment boundaries which are actual segment boundaries" divided by "the number of the true segment boundaries*". Beeferman's metric *Pk* measures the *proportion of "sentences which are wrongly predicted to belong to different segments (while they actually belong in the same segment)" or "sentences which are wrongly predicted to belong to the same segment (while they actually belong in different segments)*". A variation of the *Pk* measure named WindowDiff index was proposed by Pevzer and Hearst (Pevzer and Hearst, 2002) and remedies several of *Pk*'s problems.

It should be noted that stop word removal and stemming (i.e., substitution of a word by its root form) were performed based on Porter's algorithm (Porter, 1980) before applying the algorithms in the corpora. Table 3 contains the results reported in the literature in the original corpus as well as those obtained in the "annotated" Choi's corpus.

Table 3: Performance of three segmentation algorithms applied on the original and the "annotated" Choi's corpus.

Dataset / Algo	3-11		3-5		6-8		9-11		All Files	
Corpus	Original	Annotated	Original	Annotated	Original	Annotated	Original	Annotated	Original	Annotated
C99b Precision	78%	81.8%	85.6%	89.7%	80.7%	85.6%	86.5%	86.2%	80.7%	84.1%
Utiyama Precision	67.4%	79.4%	77.8%	82.2%	77.8%	90.6%	79.3%	87.5%	72.0%	82.6%
Kehagias Precision	82.6%	72.6%	82.1%	83.9%	88.6%	89.2%	93.3%	87.7%	85.6%	78.7%
C99b Recall	78.0%	81.8%	85.6%	89.7%	80.7%	85.6%	86.5%	86.2%	80.7%	84.1%
Utiyama Recall	70.6%	74.5%	74.2%	79.6%	86.7%	90.6%	87.7%	87.1%	75.8%	79.3%
Kehagias Recall	82.7%	70.8%	87.7%	81.7%	88.7%	89.1%	92.4%	87.7%	85.7%	77.4%
C99b Pk	12.1%	10.8%	10.4%	8.6%	9.6%	8.4%	8.5%	8.1%	11.%	9.8%
Utiyama Pk	10%	11.5%	9%	8.2%	7%	2.4%	5%	3.3%	9%	8.4%
Kehagias Pk	7%	11.7%	5.4%	7%	3%	2.6%	1.3%	1.7%	5.4%	8.3%

We reach the following conclusions based on the obtained results. Regarding Choi's C99b algorithm, a significant improvement was obtained in all measures and for all datasets. The same observation holds for the results obtained after applying the algorithm of Utiyama and Isahara, especially in datasets 6-8 and 9-11. However, Kehagias algorithm fails to obtain better performance in the first two datasets. On the contrary, in datasets 6-8 and 9-11 the difference in the -already high- performance is marginal. This is an indication that the algorithm performs better when the segment's length is high and the deviation from the expected segment length is small. The greater difference is observed in datasets 6-9 and 9-11 for all algorithms. This is justified by the fact that, in those datasets the number of named entity instances and those resulting after co-reference resolution is higher than the equivalent in the remaining ones. It must be stressed that, co-reference resolution contributed significantly to the increase of the number of entity instances per segment.

We also draw attention to the fact that, the type of named entity instance acts indirectly as a discriminative factor in the segmentation process. This is in contrast with information extraction, where the learning process takes into account the type of named entities occurring in a passage of text.

Finally, we performed manual annotation (i.e., named entity recognition and co-reference resolution) in the Stargerzers document introduced by Hearst (Hearst, 1997) using the same types of named entities. Both documents (i.e., original and "annotated") were evaluated using Chois C99b and Utiyama and Isahara algorithms. This is because they do not require training. It must be stressed that, no "official" (i.e., widely accepted) segmentation exists for this document. The application of the C99b algorithm, in both the original and "annotated" form of the document, produced exactly the same segmentation. This segmentation is not in align with the one proposed by Hearst. On the other hand, the application of Utiyama and Isahara's algorithm (in both versions of the document), produced almost the same segmentation. The only difference noticed was in the number of paragraphs contained in the last two segments among the seven produced. The latter segmentation is closer to the one proposed by Hearst. The aforementioned experiments proved that the annotation process does not falsify the segmentation outcome.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

5 CONCLUSIONS

In this paper we evaluated the benefit of incorporating information extraction techniques to enhance the performance of text segmentation algorithms. More specifically, we performed manual named entity recognition and co-reference resolution on the Choi's benchmark used by text segmentation algorithms. We then compared the performance of three well-known segmentation algorithms in both the original and the resulting "annotated" corpus. The results obtained show that, this type of annotation has an added value as the segment length increases. The potential benefit of the annotation is strongly related to the segment's topic as well as the number of named entity instances appearing in it. This approach may further prove beneficial for other problems, such as web mining and focused crawling.

We outlook several directions of future work. The first direction considers performing text segmentation on a different corpus with fewer topics than Choi's corpus, such as the Reuters RCV1 and RCV2 corpora. In these corpora named entity recognition and co-reference resolution would be performed. The second direction is oriented towards the application of named entity recognition and co-reference resolution tools in order to compare their impact in the performance of segmentation algorithms. We further seek to examine the addition of other types of named entities that will be more

oriented to the segment's topic. In the same direction lies the extraction and annotation of relations between named entities and the examination of their contribution to the segmentation task. The aim is to reinforce the role and identity of named entities in the segmentation process. Finally, it is interesting to examine the impact of named entity recognition and co-reference resolution in corpora written in other languages than English like Greek. An example of a Greek corpus used for text segmentation is the one presented in (Fragkou et al., 2007).

REFERENCES

Beeferman, D., Berger, A. and Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning,* 34:177-210.

Bestgen, Y. (2006). Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings Deterministic and Moore (2001). *Computational Linguistics,* 1:5-12.

Choi, F.Y.Y. (2000). Advances in domain independent linear text segmentation. *In Proc. of the 1st Meeting of the North American Chapter of the ACL*, pages 26-33.

Choi, F.Y.Y., Wiemer-Hastings, P. & Moore, J. (2001). Latent semantic analysis for text segmentation. *In Proceedings of the 6th Conf. on EMNLP*, pages 109 - 117.

Corcho O. (2006). Ontology based document annotation: trends and open research problems. *Int. J. Metadata, Semantics and Ontologies,* 1(1):47-57.

Fragkou, P., Petridis, V. and Kehagias, A. (2007). Segmentation of Greek Text by Dynamic Programming. In *Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007),* vol. 2, pages 370-373.

Fragkou, P. (2009). A comparison of Information Extraction and Text Segmentation for Web Content Mining. In *4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2009),* pages 482-486.

Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, *Computational Linguistics*, 23(1):33-64.

Heinonen, O. (1998). Optimal Multi-Paragraph Text Segmentation by Dynamic Programming. In *Proc. of 17th COLING -ACL'98*, pages 1484-1486.

Kehagias, Ath., Nicolaou A., Fragkou P. and Petridis V. (2004). Text Segmentation by Product Partition Models and Dynamic Programming. *Mathematical & Computer Modelling, 39:*209-217.

Kern, R. and Granitzer, M. (2009). Efficient linear text segmentation based on information retrieval techniques. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*.

Müller, C. and Strube, M. (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In *S. Braun, K. Kohn and J. Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. *English Corpus Linguistics*, 3: 197-214.

Ogren, P. V. (2006). Knowtator: A Protégé plug-in for annotated corpus construction. *Human Language Technology Conference Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273-275.

Porter, M.F. (1980). An algorithm for suffix stripping *Program*, 14(3): 130-137.

Petasis, G., Karkaletsis, V., Paliouras, G., Spyropoulos, C. D. (2003). Using the Ellogon Natural Language Engineering Infrastructure. *In Proceedings of the Workshop on Balkan Language Resources and Tools, 1st Balkan Conference in Informatics (BCI 2003)*.

Peyzner, L. and Hearst, M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics,* 28(1):19-36.

Ponte, J. M. and Croft, W. B. (1997). Text segmentation by topic. In *Proc. of the 1st Europ. Conf. on Research and Advanced Technology for Digital Libraries*, pages 120 - 129.

Qi S., Runxin L., Dingsheng L. and Xihong W. (2008). Text segmentation with LDA-based Fisher kernel. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologie*s, pages 269-272.

Reynar, J.C. (1994). An automatic method of finding topic boundaries. In *Proc. of the 32nd Annual Meeting of the ACl*, pages 331-333.

Sitbon, L. and Bellot, P. (2005). Segmentation thématique par chaînes lexicales pondérées. In *Proc the 12th Conference on Natural Language Processing (TALN 2005)*.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain independent text segmentation*. In Proc. of the 9th EACL*, pages 491-498.

Xiang J. and Hongyuan Z. (2003). Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. In *Proc. of the 26th ACM SIGIR Conf.*

Yaari, Y. (1999). *Intelligent exploration of expository texts*. Ph.D. thesis. Bar-Ilan University.

Ye, N., Zhu, J., Luo, H.,Wang, H. and Zhang, B. (2005). Improvement of the dotplotting method for linear text segmentation. *In Proc of Natural Language Processing and Knowledge Engin*eering, pages 636-641.