ORIGINAL RESEARCH

Combining information extraction and text segmentation methods in Greek texts

Pavlina Fragkou*

Department of Informatics Systems, Technological Educational Institution of Athens (TEI-A), Athens, Greece

Received: September 12, 2017 **Accepted:** December 26, 2017 **Online Published:** January 8, 2018
DOI: 10.5430/air.v7n1p23 **URL:** http://dx.doi.org/10.5430/air.v7n1p23

ABSTRACT

This paper leverages semantic information that is elicited from information extraction techniques, to text segmentation algorithms. The purpose here is to examine whether semantic information boosts segmentation accuracy. Present study is performed in a Greek corpus. Semantic extraction is performed through an already existing NER tool for Greek (focusing on four named entity types) as well as (manually performed) co-reference resolution. Produced results reveal that, the proposed approach can be very promising in improving text segmentation performance as a result of extracting valuable semantic information. They also reveal that, manual annotation in specific information extraction tasks constitutes a unique option due to lack of freely available automatic annotation tools especially in languages such as Greek.

Key Words: Text segmentation, Information extraction, Semantic search, Machine learning, Computational linguistics

1. INTRODUCTION

According to Wikipedia "*Semantic search seeks to improve search accuracy by understanding the searcher's intent and the contextual meaning of terms as they appear in the searchable dataspace, whether on the Web or within a closed system, to generate more relevant results. Semantic search systems consider various points including context of search, location, intent, variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results*." The problem thus is divided into two sub-problems: the first one lies in the identification of the aforementioned "various points" within a text passage, while the second in the identification of the location of portions of relevant information within a text.

Information extraction techniques, which deal with the first sub-problem, admit that documents referring to a specific topic describe one or more (named) entities or events in a similar manner. For example in a passage about politics, facts about the visit of a prime minister of country X to another country Y to participate in a meeting, such as his name, the location names, the date that the event took place, as well as the event's name.[1,2]

In order to reveal instances of specific types of named entities (such as person, organization, date, and location) in a form that is more closely to semantic metadata, among others, two steps are taking place. The first step is *Named entity recognition* i.e. recognition of all mentions of predefined entity names (for people and organizations), place names, temporal expressions etc, according to a specific domain and assignment of a unique identifier to each extracted entity. The second step is *co-reference* resolution, which involves the creation of anaphoric links between previously

*Correspondence: Pavlina Fragkou; Email: pfragkou@teiath.gr; Address: Department of Informatics Systems, Technological Educational Institution of Athens (TEI-A), Ag. Spyridonos, 12210, Egaleo, Athens, Greece.

identified-extracted mentions of text entities, through the detection of a number of types of links. Those mentions are revealed by personal pronouns, possessive adjectives, possessive pronouns, reflexive pronouns, pronouns '*this*' and '*that*', ordinal anaphora (for cardinal numbers and adjectives such as '*former*' and '*latter*') etc. For example, in the sentences "*Macron ends Greek visit with surprise stroll through town. French President Emmanuel Macron and his wife Brigitte Macron take a selfie with passers-by on main commercial Ermou Street in Athens, Friday*." It would be beneficial to detect that "*his*" is referring to the previously detected person "*Emmanuel Macron*". Moreover, a unique named identifier must be attributed to "*Macron*", "*Emmanuel Macron*" and "*French President*" mentions in the text.

Work regarding NER for Greek, focuses on hand-crafted rules or patterns[1,3,4] and/or decision tree induction with C4.5.[5,6] Different approach is followed by Diamantaras et al.[7] and Michailidis et al.,[8] where SVMs, Maximum Entropy, Onetime, and manually crafted post-editing rules are employed. The interest is concentrated in two works. The first work introduces an Information extraction pipeline which consists of a tokenizer, a POS tagger, and a lemmatizer. This pipeline contains tools that are able to recognize named entities, recursive syntactic structures, grammatical relations, and co-referential links by paying special attention to pronominal anaphora resolution.[9] During the second work,[10] a freely available named entity recognizer for Greek texts is constructed which is capable of identifying temporal expressions, person names as well as organization names.

Co-reference resolution is also applied as a subsequent step of NER for Greek. More specifically, Papageorgiou et al.[9] chose to focus on pronominal anaphora resolution. In their work, co-reference resolution focused only on the third person possessive and the relative pronoun. They examined both *intra*-sentential anaphora (i.e., where anaphoric links appear within the same sentence) as well as *inter*-sentential anaphora (i.e., in which the pronoun refers to an entity mentioned in a previous sentence).

The second sub-problem is dealt by information retrieval and more specifically text segmentation methods. Text segmentation's goal is the division of a document into meaningful units, such as words, sentences, or topics, each of which corresponds to a particular subject. Text segmentation methods, according to the approach followed to detect those meaningful units, can be classified as:[11] (a) *Similarity based methods*, which measure proximity between sentences. A common criterion used here is the cosine of the angle between vectors[12–15] -vectors are based on word distribution of sentences but not on named entity instances and anaphoric links resulted from co-reference resolution highlighting thus, the appearance of specific words in the scope of a particular topic; (b) *Graphical methods*, which graphically represent term frequencies and use of these representations to identify topical segments, where the most common approach is the dot plotting algorithm;[16] (c) *Lexical chains based methods*, which link multiple occurrences of a term.[17]

Moreover, additional methods exist which focus on other issues such as: (a) deviations from the expected segment length, based on the calculation of '*length model*';[15] (b) calculation of the globally minimal segmentation cost usually by using dynamic programming;[15,18–20] (c) Latent Semantic Analysis;[21] (d) the notion of "tiles" introduced by Hearst[14] as well as its improvement presented by Kern and Granitzer;[22] (e) the adoption of the Markovian assumption;[18] (f) other approaches to segment unit representation i.e., distribution of topics instead of a set of word tokens.[23]

Segmentation of Greek texts is limited dealt in the literature.[24] However, it constitutes a difficult problem due to the high degree of inflection that Greek language presents, compared to English, making thus segmentation task even harder. More specifically in English, the function that a noun performs is based upon its position in the sentence. Greek language, however, operates in a different manner where, each Greek word actually changes form based upon the role that it plays in the sentence. Verbs also change forms to indicate things such as person, tense, mood, etc. Thus, a non-trivial task consists in understanding what the case of each Greek noun indicates and what the form of every Greek verb is implying. In Greek, nouns, adjectives and verbs are each divided into several inflectional classes which have different sets of endings.

As previously mentioned, semantic search aims at penetrating to context thus, to provide meaningful units of text that represent the most accurate answer to a user's query. However, methods followed (such as Latent Semantic Indexing, Latent Dirichlet Allocation as well as TF-IDF weighting) fail to capture both semantic correlation as well as portion identification of user's answer within a text. Those issues are partially dealt by information extraction techniques and text segmentation algorithms. The paper's intention is to enhance text segmentation algorithms with semantic information extracted using information extraction techniques. The basis of the present study is the fact that segmentation algorithms do not exploit the contextual meaning of terms with regard to content in which those terms appear to. More specifically, in this paper the interest is on examining whether contextual meaning of terms, extracted either manually or automatically (by using publicly available annotation tools) succeeds in

improving segmentation accuracy. This extraction also evaluates manual annotation effort against the adequacy as well as the adaptation effort of publicly available annotation tools.

The reverse problem of the one examined here was dealt by Fragkou.[25] Special attention must be given to the fact that, the aim of the present study in on the exploitation of conceptual meaning of terms using a limited number of text segmentation algorithms applied to a Greek corpus and not on identifying the algorithm that succeeds in achieving the best segmentation accuracy.

A previous approach which presents a resemblance to the work presented here was performed by Sitbon et Bellot.[17] This work used named entity instances belonging to three types (i.e. person name, location, and organization) as lexical chains elements. In order to capture anaphoric links, the authors claim use of anaphors without however providing further implementation details. Their experiments were performed on two French corpora. According to the authors, results obtained using named entity instances fail to improve segmentation accuracy. The explanations provided by them were: (a) frequent use of anaphora can be the cause of restriction of named entity repetition; (b) restriction of the number of features used may be attributed to the use of lexical chains.

The novelty of current work lies in the incorporation of semantic information produced after applying NER and (manual) co-reference resolution (including all types of anaphora resolution) to text segmentation algorithms. Emphasis must be given to the fact that, research for segmenting Greek texts focuses on segmenting ancient manuscripts using optical character recognition techniques. Even though other datasets containing (modern) text exist, those were subjected to other studies than text segmentation. Moreover, no other researchers have dealt the problem of text segmentation in the way that is approached in the current study. This prevents additional evaluations and comparisons with other researchers for validating obtained results.

The improvement of present work against the one presented by Sitbon et Bellot[17] lies in the following factors: (a) the improvement achieved after restricting the scope of manual co-reference resolution to those terms that are related only with named entity instances; (b) the comparison of manual annotation with the output produced using a publicly available automated NER annotation tool; (c) the exploitation of another frequently used named entity type i.e., date; (d) the assessment of the quality of produced annotated corpus using four text segmentation algorithms.

The work presented here is organized as follows: Section 2 deals with the creation of the 'annotated' corpus for Greek. Section 3 focused on the text segmentation algorithms used for experiments, while Section 4 provides the evaluation metrics used to calculate segmentation performance. Section 5 focuses on the experiments performed as well as obtained results after applying four widely used segmentation algorithms to the Greek corpus, while Section 6 lists conclusions and future steps.

2. GREEK CORPUS

Publicly available NER tools are extensively examined in recent years with regard to their effectiveness and appropriateness for a specific task. However, the majority of studies focus in English.[26–30] A common conclusion to all studies is that readily - available tools not only are inevitably subjected to training but they are able to identify a unique or a restricted number of topics.

Open issues regarding the effectiveness and validity of those tools - as far as the named entity annotation outcome is concerned - are: (a) potential need for further (manual) correction and/or enhancement; (b) the scope of NER covered, in other words, the quality of named entity types in terms of their compliance with the topic in question and their nature (too generic or too specific); (c) the scope of co- reference resolution covered (i.e., within the same and/or adjacent sentences) and types of anaphora covered; (d) the ability to identify and tag appropriately all mentions (resulted from both NER and co-reference resolution) of the same named entity instance; (e) potential post-processing of the annotation outcome in order to be appropriately processed by a text segmentation algorithm.

It can be seen that, regardless language, use of already available information extraction tools involves a number of parameters:

(1) Finding a correct tool or combination of tools already trained with the most thematically related topic(s) to achieve high accuracy in recognizing named entity instances including mentions resulting from co-reference resolution.
(2) Evaluating the characteristics of annotation tools used. This means that statistical distribution of terms (i.e., words and named entity instances identified and extracted from tools) is strongly affected by the number of named entity instances captured as well as their distribution to entity types (i.e., the risk of attributing a significant number of named entity mentions to the default named entity type) which have a significant impact in segmentation accuracy.
(3) Produced annotations from NER and/or co-reference resolution tool(s) used may require manual correction

and completion of the outcome for the needs of the problem in question.

For experiments, the corpus created in[24] was used. There, the authors used a text collection compiled from Stamatatos Corpus,[31] comprising of text downloaded from newspaper's website 'To Vima'. Stamatatos et al.[31] constructed a corpus collecting texts from ten different authors. Thirty texts were selected from each author. Table 1 lists the authors contributing to Stamatatos corpus as well as the thematic area(s) covered by each of them.

Table 1. List of authors and thematic areas for Stamatatos Corpus as well as statistics regarding the average number of named entity instances in the annotated documents of the corpus per author.

Author	Thematic Area	Average number of NEs
Alachiotis	Biology	44.00
Babiniotis	Linguistics	70.23
Dertilis	History, Society	33.33
Kiosse	Archeology	121.90
Liakos	History, Society	77.70
Maronitis	Culture, Society	40.40
Ploritis	Culture, History	94.20
Tassios	Technology, Society	40.00
Tsukalas	International affairs	37.12
Vokos	Philosophy	52.16

The corpus used here is the one that was created in.[24] Its articles were subjected to POS tagging using Orphanos and Christodoulakis tool.[32] At a subsequent step, a selection of specific types of words i.e. nouns, verbs, adjectives and adverbs, was performed. For each selected type, its lemma as determined by the tagger was chosen. The previously described corpus was also used for current experiments due to its uniqueness to the problem examined.

As it was pointed out in Section 1, the number of readily-available automated annotation tools for Greek is very low. For current experiments, the automated annotation tool described in Lucarelli et al.[10] in the corpus created in[24] was applied. This annotation tool was chosen because it is publicly available, it is trained on similar documents taken from newspaper 'Ta Nea' and produces an output that can be easily processed by a text segmentation algorithm. Newspaper 'Ta Nea' contains articles having similar content with that of newspaper 'To Vima'. The annotation tool was thus applied in the corpus without requiring training. Four named entity types were chosen i.e., person name, group name, location, and date. The annotation tool produced annotations for some, but not all instances of person names, group names, and dates. In order to annotate all named entities appearing in each text, a second pass was performed. During this pass, enhancement of named entity instances (restricted to proper names belonging to the four preselected named entity types) was performed manually in each article. No correction was performed, since the annotation tool was proven to perform correct annotation to those named entity instances that could identify and appropriately classify. During manual completion of named entity annotation: (a) all instances of locations were additionally annotated and (b) the same named entity identifier was attributed to all references of the same instance i.e., co-reference resolution was performed to identify all mentions that represent the same entity and grouping of them to the entity they refer to, by paying special attention to the appearance of Greek pronouns. Co-reference resolution was performed manually concentrating on portions of text that are associated with named entity instances by identifying possessive and personal pronouns as well as pronominal anaphora. The latter step was necessary because the annotation tool used cannot perform co-reference resolution. It must be stressed that, no parser was needed to be constructed since produced output was in a form that could be easily processed by a segmentation algorithm.

Table 2 provides a portion of a document belonging to Tsulakas author, in its original form, after applying Orphanos POS tagger and subsequently performing stop list removal, after applying Lucarelli's et al. annotation tool, as well as subsequently performing manual annotation (including co-reference resolution step). Underlined and bolded words in the annotated corpus correspond to named entity instances. This portion of text may constitute a portion of text in a document processed by a text segmentation algorithm.

The annotation process led to the conclusion that, texts having a social subject usually contain a small number of named entity instances, contrary to texts about politics, science, archeology, history, and philosophy. For example, texts belonging to the author Kiosse contain on average large number of named entities, because they describe historical events issuing person names, dates, and locations. Table 1 provides lists for every author in the corpus, the average number of named entity instances appearing in its annotated documents.

3. SEGMENTATION ALGORITHMS

Four text segmentation algorithms were selected to evaluate segmentation accuracy obtained by the annotated corpora that resulted from the application of a named entity annotation tool (only) and the substitution of each entity mention by a unique named entity identifier, as a result of restricted co-reference resolution.

The first algorithm is Choi's C99b[12, 13] which uses a matrix-based ranking applied on a cosine similarity matrix of sen-

tences and a top-down hierarchical clustering approach, in order to relate the most similar textual units and to cluster groups of consecutive units into segments.

Table 2. Portion of a document of the Greek corpus belonging to Tsukalas author, before performing NER and co-reference resolution, after applying Lucarelli's et al. automated annotation tool, as well as performing manual co-reference resolution.

Original Text	<CC> Κ. ΤΣΟΥΚΑΛΑΣ ΤΟ ΒΗΜΑ, 06-06-1999 Κωδικός άρθρου: Β12598Α381</CC> <TITLE> Οι Γάλλοι και οι άλλοι Η Γαλλία κινδυνεύει να αποδεχθεί την αποδυνάμωση του πολιτισμού της </TITLE> <TEXT>Μόλις γύρισα από τη Γαλλία, τη χώρα που με φιλοξένησε μαζί με τόσους άλλους που δεν μπορούσαν ή δεν ήθελαν να γυρίσουν στη χουντοκρατούμενη Ελλάδα, τη χώρα από την οποία κατ' εξοχήν έλκω την πνευματική κατάρτισή μου, τη χώρα που αγάπησα όσο καμία άλλη εκτός από τη δική μου. Και γύρισα με κόμπο στην καρδιά και αμφιβολία στην ψυχή. Η Γαλλία δεν είναι πια αυτή που ήξερα ή εκείνη που νόμιζα ότι ξέρω.
Orphanos POS Tagger, Lemmatization, stop list removal	<SENTENCE> Κ. ΤΣΟΥΚΑΛΑΣ ΤΟ ΒΗΜΑ, 06-06-1999 Κωδικός άρθρο: Β12598Α381</ SENTENCE > <SENTENCE > Γάλλοι Γαλλία κινδυνεύω να αποδεικνύομα αποδυνάμωση πολιτισμός </ SENTENCE ><SENTENCE> γυρίζω Γαλλία χώρα φιλοξενώ μπορώ θέλω γυρίζω χουντοκρατούμενη Ελλάδα χώρα εξοχή έλκω πνευματικός κατάρτιση χώρα αγαπώ </SENTENCE> <SENTENCE> γυρίζω κόμπος καρδιά αμφιβολία ψυχή </SENTENCE> <SENTENCE> Γαλλία ξέρω νομίζω ξέρω </SENTENCE>
Luccarelli NER	<ARTICLE> <SENTENCE><ENAMEX TYPE="PERSON" CONF0="0.20294629407758313"> Κ</ENAMEX>. ΤΣΟΥΚΑΛΑΣ ΤΟ ΒΗΜΑ, 06-06-<TIMEX TYPE="DATE">1999</TIMEX> Κωδικός άρθρου: Β12598Α381 Οι Γάλλοι και οι άλλοι Η Γαλλία κινδυνεύει να αποδεχθεί την αποδυνάμωση του πολιτισμού της. Μόλις γύρισα από τη Γαλλία, τη χώρα που με φιλοξένησε μαζί με τόσους άλλους που δεν μπορούσαν ή δεν ήθελαν να γυρίσουν στη χουντοκρατούμενη Ελλάδα, τη χώρα από την οποία κατ' εξοχήν έλκω την πνευματική κατάρτισή μου, τη χώρα που αγάπησα όσο καμία άλλη εκτός από τη δική μου.</SENTENCE><SENTENCE>Και γύρισα με κόμπο στην καρδιά και αμφιβολία στην ψυχή.</SENTENCE><SENTENCE>Η Γαλλία δεν είναι πια αυτή που ήξερα ή εκείνη που νόμιζα ότι ξέρω.</SENTENCE> <SENTENCE>
Manual Annotation (co-reference resolution and attribution of a unique NE identifier)	Άτομοένακείμενοεικοσιένασι γκρούπένακείμενοεικοσιένασι, χρονολογίακείμενοεικοσιένασι Κωδικός άρθρου:Β12598Α381 Οι τοποθεσίαένακείμενοεικοσιένασι και οι άλλοι Η τοποθεσίαένακείμενοεικοσιένασι κινδυνεύει να αποδεχθεί την αποδυνάμωση του πολιτισμού της τοποθεσίαένακείμενοεικοσιένασι. Μόλις γύρισα από τη τοποθεσίαένακείμενοεικοσιένασι , τη χώρα που με άτομοένακείμενοεικοσιένασι φιλοξένησε μαζί με τόσους άλλους που δεν μπορούσαν ή δεν ήθελαν να γυρίσουν στη χουντοκρατούμενη τοποθεσίατρίακείμενοεικοσιένασι, τη χώρα τοποθεσίαένακείμενοεικοσιένασι από την οποία τοποθεσίαένακείμενοεικοσιένασι κατ' εξοχήν έλκω την πνευματική κατάρτισή μου άτομοένακείμενοεικοσιένασι, τη χώρα τοποθεσίαένακείμενοεικοσιένασι που αγάπησα όσο καμία άλλη εκτός από τη δική μου άτομοένακείμενοεικοσιένασι. Και γύρισα με κόμπο στην καρδιά και αμφιβολία στην ψυχή. Η τοποθεσίαένακείμενοεικοσιένασι δεν είναι πια αυτή που ήξερα ή εκείνη που νόμιζα ότι ξέρω.

The second algorithm introduced by Utiyama and Isahara was also chosen.[33] This algorithm calculates the probability of words belonging to a segment via a statistical model. To accomplish this, the algorithm seeks to maximize the probability of a segmentation S given a word sequence W. The algorithm uses maximum – likelihood estimation and Laplace smoothing to compute the parameters of a dynamic programming algorithm. The idea is that each topic is characterized by a word distribution. Software for Choi's C99 as well as Utiyama and Isahara's algorithm is publicly available and can be applied without requiring training.

This does not hold for the third algorithm used here (Kehagias et al.[34]) which requires training. This algorithm treats segmentation as an optimization problem with global cost function and uses dynamic programming for segments selection i.e., the algorithm identifies the number of segments along with their position within the text. The algorithm performs linear segmentation by minimizing the global segmentation cost which is based in a similarity matrix, a preferred fragment length, and a cost function defined.

The last algorithm named Affinity Propagation considered four our experiments was implemented by Kazantseva and Szpakowicz.[18] Its version for the segmentation task takes as input a set of pairwise similarities between data points. Segment boundaries as well as segment centres i.e., data points/ clusters which best describe all other data points within the segment are identified by iteratively passing messages in a cyclic factor graph which considers all available similarities, until convergence. The preference parameter is the value which clusters the N data points into N clusters, and this is equal to the maximum similarity.

4. EVALUATION METRICS

The performance of the algorithms applied in the annotated corpora was calculated using four widely known metrics: Precision, Recall, Beeferman's Pk[35] and WindowDiff.[36] For the segmentation task, Precision is defined as the proportion of boundaries chosen that agree with a reference segmenta-

tion. In a similar manner, Recall is defined as the proportion of boundaries chosen that agree with a reference segmentation out of all boundaries in the reference and hypothesis. However, both metrics suffer by the fact that they penalize near-misses of boundaries as full-misses, causing them to drastically overestimate the error. Beeferman's Pk[35] metric attempts to correct the erroneous calculation of penalties performed by Precision and Recall by computing penalties using a sliding a window of size k across the text, where k is defined as half of the mean reference segment size. Penalties are calculated by taking into account both the number of windows as well as whether boundaries appear in different segments in the reference and in the hypothesis segmentations for every window examined. It must be stressed that Beeferman's Pk metric measures segmentation inaccuracy. Thus high segmentation accuracy is achieved when small values of the metric are obtained.

WindowDiff metric proposed by Pevzner and Hearst[36] is a variant of the Pk measure, which penalizes false positives and near misses equally. WindowDiff also follows a window based approach but errors are associated with windows (a window is evaluated as either correct or incorrect). WindowDiff is sensitive to the balance of positive and negative data being evaluated and consequently to window size. Even thought, other evaluation metrics are recently proposed,[37,38] the aforementioned evaluation metrics were chosen since they are the most widely used.

5. Experiments

The effect in segmentation accuracy of an algorithm after replacing every word or phrase with a unique named entity instance is examined here, using the algorithms presented in Section 3 and the four evaluation metrics defined in Section 4. More specifically, two groups of experiments using the corpus presented in Section 2 were performed. Figure 1 presents all steps followed by the approach analyzed here. For both groups, three different versions of the corpus are considered (i.e., the original/non-annotated corpus, the annotated corpus using Lucarelli's et al. annotation tool and the manual annotated corpus) which are subjected to segmentation accuracy comparison.

5.1 First group of experiments

For the first group of experiments Greek corpus examined in[24] was used. More specifically, six datasets: Set0,..., Set5, whose difference lies in the number of authors contributing in the generation of the texts to segment, thus the number of texts originated from the entire collection were created.[31] The first dataset use documents taken from authors Kiosse and Alachiotis, the second from authors Kiosse and Maronitis, the third from authors Kiosse, Maronitis and Alachiotis, the fourth from authors Kiosse, Maronitis, Alachiotis and Ploritis, the fifth from authors Kiosse, Maronitis, Alachiotis, Ploritis and Vokos, while the sixth from all ten authors.

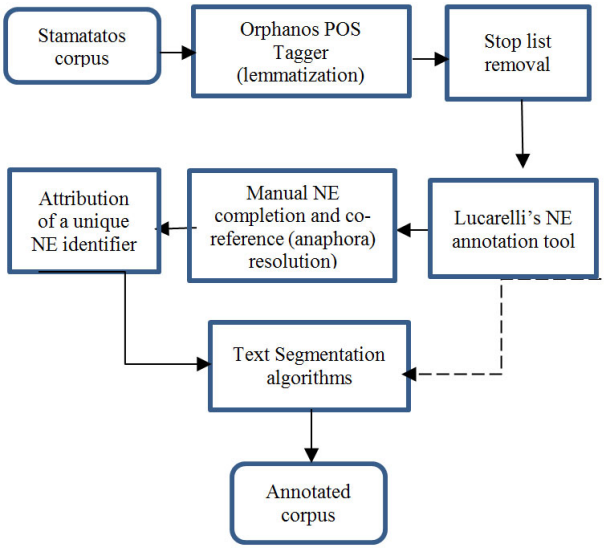

Figure 1. Flowchart describing our approach i.e. the annotation step - by performing automatic NER using Lucarelli's NE annotation tool and manual co-reference - as well as the segmentation step.

For each of the above datasets, four subsets were constructed, each of which differs in the number of sentences appearing in each segment. More specifically, each subset belongs to one of the pairs (3,11); (3,5); (6,8); and (9,11) where the first element in the pair corresponds to the smallest number of sentences that a segment may contain while the second element to the largest one. The notation Set*1 to denote all datasets belonging to pair (3,11), Set*2 all datasets belonging to pair (3,5), and so on was used.

Each subset contains fifty documents to segment, each of which is a concatenation of ten segments. Each segment belongs to a portion of a randomly selected document (among the thirty available) of a randomly selected author contributing in this dataset. This portion contains a randomly selected sentences of the author's document (starting from the first sentence of the text) taking into account the restriction imposed by the subset in which the document to segment belongs to.

Segmentation accuracy was measured using all four algorithms in all datasets. Affinity Propagation fails to produce results for both the annotated and non-annotated corpus for Set*2 (3-5). It seems that the algorithm fails to identify exemplars i.e., cluster centers for every examined document. This can be attributed to the small number of sentences appearing

in Set*2 (3-5), the choice of parameter values (especially the 'preference' parameter) as well as problems encountered with similarity measurement yielding to oscillations i.e., no convergence. This is related to the well-known disadvantages of Affinity Propagation i.e., (a) it is hard to know the value of "preference" parameter which can yield an optimal clustering solution; (b) the "preference" parameter is self-adaptive mainly according to the target number of clusters. Thus, in Set*2 (3-5) where the number of sentences in a segment varies from three to five, the algorithm fails to identify appropriate clusters i.e. to find a good value which can optimize the clustering result. Problems regarding similarity measurement as well as parameter selection are addressed in subsequent implementations of the algorithm.

Obtained results after applying each of the four segmentation algorithms measured using each of the four evaluation metrics: (a) in the non-annotated corpus; (b) in the manual annotated corpus; (c) in the corpus resulting after applying Lucarelli's et al. annotation tool only, averaged over all datasets which have segments of the same length are listed in Table 3. Italic notation is used to denote the average performance obtained by each segmentation algorithm over all datasets.

Table 3. Precision, Recall, Beeferman's Pk and WindowDiff values (percent) obtained by the four algorithms in the first group of experiments without and with use of named entities for Greek texts as well as use of Lucarelli's NE annotation tool only

Algorithm	Dataset	Precision No Annotation	Precision NEs Annotation	Precision with Annotation	Recall No Annotation	Recall NEs Annotation	Recall with Annotation	Pk No Annotation	Pk NEs Annotation	Pk With Annotation	Window Diff No Annotation	Window Diff NEs Annotation	Window Diff With Annotation
Choi's C99b	Set*1 (3-11)	59.7	48.73	63.26	59.67	44.51	63.26	17.96	19.54	15.96	19.37	21.34	17.40
	Set*2 (3-5)	67.86	52.9	70.46	67.86	46.94	70.46	16.70	20.82	14.53	17.93	22.93	15.91
	Set*3 (6-8)	64.9	54.57	71.26	64.9	54.56	71.26	15.13	14.06	11.92	15.89	15.17	12.45
	Set*4 (9-11)	64.23	53.1	68.46	64.23	48.8	68.46	13.60	16.30	11.43	14.07	16.71	11.89
	All Files	*64.17*	*52.32*	*68.36*	*64.17*	*48.70*	*68.36*	*15.85*	*17.68*	*13.46*	*16.82*	*19.04*	*14.41*
Utiyama & Isahara	Set*1 (3-11)	64.18	51.69	70.74	61.24	44.99	66.96	17.47	18.38	13.72	18.48	19.50	14.70
	Set*2 (3-5)	70.04	54.37	76.65	54.74	35.68	61.55	20.99	26.24	16.83	21.31	26.71	17.30
	Set*3 (6-8)	75.45	55.45	80.31	73.07	52.74	78.18	10.96	14.00	8.43	11.00	14.07	8.44
	Set*4 (9-11)	73.17	56.98	76.75	74.33	57.25	78.40	8.91	9.70	7.15	9.03	9.86	7.34
	All Files	*70.71*	*54.62*	*76.11*	*65.84*	*47.67*	*71.27*	*14.58*	*17.08*	*11.53*	*14.95*	*17.53*	*11.95*
Kehagias et al.	Set*1 (3-11)	64.90	53.88	70.12	61.77	51.04	67.92	15.69	15.05	13.12	17.16	17.208	14.67
	Set*2 (3-5)	85.13	62.94	87.58	85.11	49.15	87.48	6.45	9.16	5.15	6.52	13.68	5.24
	Set*3 (6-8)	90.51	80.51	92.29	90.51	80.51	92.29	2.54	2.76	2.04	2.47	2.69	1.96
	Set*4 (9-11)	91.92	83.88	93.11	91.92	83.88	93.11	1.29	1.59	1.10	1.23	1.534	1.03
	All Files	*83.12*	*70.30*	*85.78*	*82.33*	*66.15*	*85.2*	*6.49*	*7.14*	*5.35*	*6.84*	*8.78*	*5.73*
Affinity Propagation	Set*1 (3-11)	24.34	15.43	24.5	17.72	11.8	17.85	-	-	-	33.31	34.08	33.29
	Set*2 (3-5)	27.41	25.53	27.41	24.67	23.44	24.67	-	-	-	N/A	5.80	N/A
	Set*3 (6-8)	22.70	29.47	22.71	22.33	29	22.33	-	-	-	31.12	31.11	31.07
	Set*4 (9-11)	15.43	20.47	15.48	15.42	20.4	15.48	-	-	-	25.55	28.01	25.42
	All Files	*22.47*	*15.43*	*22.53*	*0.03*	*11.8*	*20.08*	*-*	*-*	*-*	*22.5*	*34.08*	*22.44*

From obtained results the following conclusions can be made: Affinity Propagation algorithm's performance is -marginally- better in the annotated corpus in all datasets. Regarding Utiyama and Isahara's algorithm, a remarkable increase in performance was obtained in all measures and for all datasets of this group of experiments i.e., for manual annotation. Choi's C99b and Kehagias et al. algorithms perform similarly i.e., improvement can be observed in all evaluation metrics and for all datasets and for manual annotation. This improvement appears to be greater in datasets Set*1 (3-11) and Set*2 (3-5) in all algorithms. This is an indication that the annotation succeeded in identifying critical information which, in other ways, was lost. Segmentation precision in datasets Set*3 (6-8) and Set*4 (9-11), remains remarkable. The reason for this is that in those datasets, segment length is high leading to a high number of named entity instances.

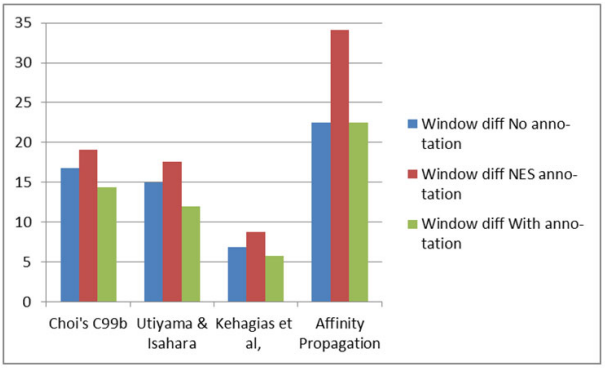

Figure 2. Performance (measured by WindowDiff) of the four segmentation algorithms applied in all datasets i.e. a) the non-annotated corpus; b) the manually annotated corpus; c) the corpus produced using Lucarelli's NER tool only

Segmentation accuracy is significantly improved after applying co-reference resolution instead of NE annotation only.

This is apparent in the performance obtained by all algorithms in all datasets (except for Precision and Recall obtained by Affinity Propagation algorithm in Set*3 (6-8) and Set*4 (9-11)). Lucarelli's annotation tool, when used alone, fails to improve segmentation accuracy compared to the one obtained in non-annotated texts. An explanation to this is that NE annotation tool is able to recognize only NE instances but not mentions of the same named entity instance. Figure 2 depicts obtained performance of segmentation algorithms measured by WindowDiff metric for all datasets.

5.2 Second group of experiments

The validity of the expectation expressed earlier was tested on a second collection which consists of 200 documents. Documents belonging to this collection also originate from Stamatatos corpus. The difference between collections lies in the fact that, each segment appearing in a document results from an arbitrary (randomly selected) number of paragraphs (and not sentences) originated from an author's document (which is also randomly selected). Those selected paragraphs may appear at any position of the author's document (not necessarily at the beginning of it). A document among the 200 created, contains portions of documents belonging to all ten authors (each segment corresponds to a different author and each document results from concatenating ten segments). The order in which an author's portion of a document appears results randomly. Consequently, the segmentation task in this collection becomes harder since, segments and consequently concatenated texts are longer in length compared to those used in the first group of experiments.

Table 4 lists values obtained by the four segmentation metrics after applying the four segmentation algorithms on the original (i.e., non-annotated) corpus, the output produced after applying Lucarelli's annotation tool only, as well as respective values after applying the same algorithms on this unique dataset, where annotation including co-reference resolution was previously performed. Table 4 reveals that segmentation performance was improved in the annotated corpus using both NE and co-reference resolution for all accuracy metrics and for all algorithms, with the exception of Window Diff performance for Utiyama and Isahara's algorithm, where a slight decrease is observed. Figure 3 depicts obtained performance of segmentation algorithms measured by WindowDiff metric in the current dataset.

Table 4. Precision, Recall, Beeferman's Pk and WindowDiff values (per cent) obtained by the four algorithms in the second group of experiments without and with use of named entities for Greek texts as well as use of Lucarelli's NE annotation tool only.

Algorithm	Precision No Annotation	Precision NEs Annotation	Precision With Annotation	Recall No Annotation	Recall NEs Annotation	Recall with Anno tation	Pk No Annotation	Pk NEs Annotation	Pk With Annotation	WindowDiff No Annotation	WindowDiff NEs Annotation	WindowDiff With Annotation
Choi's C99b	44.62	39.65	49.40	44.62	39.65	49.40	19.44	19	18.12	21.62	21	20.47
Utiyama & Isahara	56.76	47.15	59.78	67.22	55.05	69.00	12.28	11.84	10.83	12.26	14.77	13.57
Kehagias et al.	60.60	49.90	63.46	7.00	48	62.00	11.07	11	9.06	11.06	13.40	9.30
Affinity Propagation	8.83	2.82	9	14.64	5	14.91	-	-	-	57.38	57.22	57.30

The increase in segment length in this collection justifies the increase in the number of named entity instances which consequently has a positive impact in segmentation's performance. Moreover, increase in segmentation's performance also results after applying co-reference resolution since it proves to increase the number of named entity instances per segment.

Experiments using Lucarelli et al. tool only exhibit lower performance compared with the one obtained in the non-annotated corpus, except for all values of Beeferman's Pk as well as WindowDiff performance obtained by Choi's C99b and Affinity Propagation algorithms. This can be attributed to the fact that, the annotation tool augments text vocabulary with the presence of unnecessary tags. Additionally, obtained performance by the application of co-reference resolution step at the output produced by Lucarelli's et al. NER tool, proves the effectiveness and importance of both steps. This arises from obtained results in all metrics and for all algorithms, except for Affinity Propagation's WindowDiff performance, where a slight decrease is observed.

Segmentation is affected indirectly by named entity instance distribution and more specifically person name and group name. The argument for this is that, in Stamatatos corpus person names are the first most frequent named entity type and group name is the second one, since it is used as the "default" named entity type. Segmentation is also affected by named entity type selection, which must be compliant with document's topic. Appropriate selection results in cor-

rect assignment of instances to named entity types which has an impact in named entity distribution and consequently reinforces intra segment similarity.

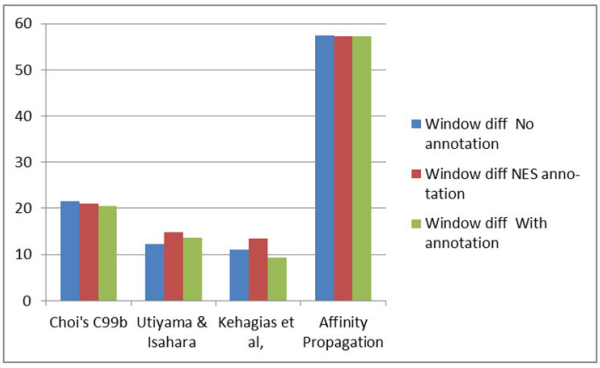

Figure 3. Performance (measured by WindowDiff) of the four segmentation algorithms applied in the dataset containing only paragraphs i.e. a) the non-annotated corpus; b) the manually annotated corpus; c) the corpus produced using Lucarelli's NER tool only.

Conducted experiments reveal that, segmentation is improved by applying co-reference resolution since already recognized mentions of named entity instances are further exploited and increased in number. The latter has an impact in named entity distribution and reinforces intra segment similarity and consequently segmentation accuracy. The way in which mentions of named entity instances are treated differs from classical information extraction, where the emphasis is on the type of named entities only but not on the value itself.

This improved performance in both groups of experiments can be attributed to the annotation performed using Lucarelli's tool,[10] which is trained on similar topics. As a consequence, no correction was required since the tool was proven to perform correct annotations to those named entity instances that could identify and appropriately classify.

6. Conclusions

This paper conducts a research regarding the contribution of semantic information attributed using information extraction techniques in the performance of text segmentation algorithms. Present study focus on a (manually constructed) Greek corpus resulted after using an automated NE annotation tool, as well as conducting manual completion of produced entity instances and mentions through manual co-reference resolution. Four segmentation algorithms were applied in the original corpus, the one resulting after applying the automated NE annotation tool only, as well as the one also containing co-reference mentions. Performance obtained by all algorithms justifies the assumption that, the more segment length increases the more named entity annotation and co-reference resolution enhances segmentation accuracy. A credit to this must be given to the appropriateness of the automated annotation tool applied. Produced results reveal that, the proposed approach can be very promising in improving text segmentation performance and consequently efficient identification of different topics that appear in documents -especially in languages such as Greek- since it reveals valuable semantic information. Obtained results are in accordance with observations made by Appelt, Atdağ and Labatut, Marrero and Siefkes.[26, 39–41]

The contribution of co-reference resolution in the improved segmentation accuracy is high and deserves special attention. The latter has an added value in languages such as Greek, which is a high inflectional language. The benefit of performing manual annotation (for co-reference resolution) instead of automatic annotation (NER only) was also examined. Emphasis must be given to the fact, automatic co-reference resolution cannot be performed for Greek since – as far as the author is aware of- the only tool appearing in the literature performing co-reference (i.e., Papageorgiou et al. tool[9]) is not publicly available.

Experiments performed in the present study, verify the assumption regarding the merit of information resulted from NER and co-reference resolution for the segmentation task. The issue here is how to obtain this valuable information. Manual annotation (for co-reference resolution) seems to be the more effective solution or the unique option due to lack of freely available automatic annotation tools. The question that arises is whether manual annotation is less time consuming than firstly choosing the adequate annotation tool(s) (from the publicly available ones), secondly conducting additional annotation and thirdly associating all mentions referring to same named entity instance with a matchless named entity identifier. Segmentation accuracy can act as yardstick of the effectiveness of each co-reference resolution type chosen.

Future work is oriented, first of all in the repetition of current experiments in a different corpus such as the one used by Lucarelli et al.[10] containing fewer topics as well as the one used by Papageorgiou et al.,[9] where co-reference resolution was also performed. The second direction is towards finding other readily - available tools and ideally tools with paying special attention in co-reference resolution (regarding its scope as well as the types of co-reference resolution covered).

Since training of a tool, consists a significant problem, interest is oriented in examining bootstrapping algorithms that can be applied to annotation tools in order to enhance training and adaptation to different topics.

Finally, since named entity types play an important role in named entity instance distribution, their examination deserves special attention. Another way to study named entity contribution in the segmentation process is by examining relations between them. A similar problem is presented in[42] where the authors proposed a two stage graphical model which first classifies entity mentions (after performing a tagging process) and then creates clusters of mentions for each distinct entity (co-reference resolution).

REFERENCES

[1] Farmakiotou D, Karkaletsis V, Koutsias J, et al. Rule-based named entity recognition for Greek financial texts. In: Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries. 2000.

[2] Grishman R. Information Extraction: Techniques and Challenges. In M.T.Pazienza (ed.) Information Extraction: A Multidisciplinary Approach to an Emergine Information Technology (International Summer School SCIE-9, Frascatti, Italy, Springer-Verlag. 1997: 10–27.

[3] Boutsis S, Demiros I, Giouli V, et al. A system for recognition of named entities in Greek. In: Proceedings of the 2nd International Conference on Natural Language Processing. 2000: 424–435.

[4] Farmakiotou D, Karkaletsis V, Samaritakis G, et al. Named entity recognition in Greek Web pages. In: Proceedings of the 2nd Hellenic Conference on Artificial Intelligence, companion volume. 2002: 91–102.

[5] Karkaletsis V, Paliouras G, Petasis G, et al. Named-entity recognition from Greek and English texts. Intelligent and Robotic Systems. 1999; 26: 123-135. https://doi.org/10.1023/A:1008124406923

[6] Petasis G, Vichot F, Wolinski F, et al. Using machine learning to maintain rule-based named-entity recognition and classification systems. In: Proceedings of the 39th Annual Meeting of ACL and 10th Conference of EACL. 2001: 426–433.

[7] Diamantaras K, Michailidis I, Vasileiadis S. A very fast and efficient linear classification algorithm. In: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing. 2005.

[8] Michailidis I, Diamantaras K, Vasileiadis S, et al. Greek named entity recognition using Support Vector Machines, Maximum Entropy and Onetime. In: Proceedings of the 5th International Conference on Language Resources and Evaluation. 2006: 45–72.

[9] Papageorgiou H, Prokopidis P, Demiros I, et al. Multi-level XML-based Corpus Annotation. In: Proceedings of the 3nd Language Resources and Evaluation Conference. 2002.

[10] Lucarelli G, Vasilakos X, Androutsopoulos I. Named Entity Recognition in Greek Texts with an Ensemble of SVMs and Active Learning. International Journal on Artificial Intelligence Tools. 2007; 16(6): 1015-1045. https://doi.org/10.1142/S0218213007003680

[11] Prince V, Labadie A. Text segmentation based on document understanding for information retrieval. In: Proceedings of the International Conference on Application of Natural Language to Information Systems NLDB 2007: Natural Language Processing and Information Systems. 2007: 295-304.

[12] Choi FYY. Advances in domain independent linear text segmentation. In: Proceedings of the 1st Meeting of the North American Chapter of the ACL. 2000: 26–33.

[13] Choi FYY, Wiemer-Hastings P, Moore J. Latent semantic analysis for text segmentation. In: Proceedings of the 6th Conference on EMNLP. 2001: 109–117.

[14] Hearst MA. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. Computational Linguistics. 1997; 23(1): 33–64.

[15] Heinonen O. Optimal Multi-Paragraph Text Segmentation by Dynamic Programming. In: Proceedings of the 17th COLING -ACL'98. 1998: 1484–1486.

[16] Ye N, Zhu J, Luo H, et al. Improvement of the dotplotting method for linear text segmentation. In: Proceedings of the Natural Language Processing and Knowledge Engineering. 2005: 636–641.

[17] Sitbon L, Bellot P. Segmentation thematique par chaines lexicales ponderees. In: Proceedings of 12th Conference on Natural Language Processing (TALN 2005). 2005.

[18] Kazantseva A, Szpakowicz S. Linear Text Segmentation Using Affinity Propagation. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011: 284–293.

[19] Qi S, Runxin L, Dingsheng L, et al. Text segmentation with LDA-based Fisher kernel. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies. 2008: 269–272.

[20] Xiang J, Hongyuan Z. Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. In: Proceedings of the 26th ACM SIGIR Conference. 2003.

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

[22] Kern R, Granitzer M. Efficient linear text segmentation based on information retrieval techniques. In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems. 2009.

[23] Yu K, Li Z, Guan G, et al. Unsupervised Text Segmentation using LDA and MCMC. In: Proceedings of the Tenth Australasian Data Mining Conference (AusDM 2012), Sydney, Australia. 2012.

[24] Fragkou P, Petridis V, Kehagias A. Segmentation of Greek Text by Dynamic Programming. In: Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence. 2007; (2): 370-373.

[25] Fragkou P. A comparison of Information Extraction and Text Segmentation for Web Content Mining. In: Proceedings of 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. 2009: 482–486.

[26] Atdağ S, Labatut V. A Comparison of Named Entity Recognition Tools Applied to Biographical Texts. In: Proceedings of the 2nd International Conference on Systems and Computer Science, Villeneuve d' Ascq (FR). 2013: 228–233.

[27] Brando C, Dominguès C, Capeyron M. Evaluation of NER systems for the recognition of place mentions in French thematic corpora. In: Proceedings of the 10th Workshop on Geographic Information Retrieval. 2016; (7): 1-7, 10.

[28] Dlugolinský S, Krammer P, Ciglan M, et al. Combining Named Entity Recognition Tools. In: Proceedings of the 3rd workshop on 'Making Sense of Microposts', (#MSM2013) World Wide Web Conference. 2013.

[29] Jiang R, Banchs RE, Li H. Evaluating and Combining Named Entity Recognition Systems. In: Proceedings of the Sixth Named Entity Workshop, joint with 54th Association for Computational Linguistics, Berlin, Germany. 2016: 21–27.

[30] Pinto A, Oliveira HG, Alves AO. Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text. In: Proceedings of the 5th Symposium on Languages, Applications and Technologies (SLATE'16), Open Access Series in Informatics (OASIcs), ISBN 978-3-95977-006-4. 2016: 1–16.

[31] Stamatatos E, Fakotakis N, Kokkinakis G. Computer-based authorship attribution without lexical measures. Computer and the Humanities, Kluwer Academic Publisher. 2001; 35: 193-214. https://doi.org/10.1023/A:1002681919510

[32] Orphanos G, Christodoulakis D. Part-of-speech disambiguation and unknown word guessing with decision trees. In: Proceedings of the EACL'99. 1999.

[33] Utiyama M, Isahara H. A statistical model for domain independent text segmentation. In: Proceedings of the 9th EACL. 2001: 491–498.

[34] Kehagias A, Nicolaou A, Fragkou P, et al. Text Segmentation by Product Partition Models and Dynamic Programming. Mathematical & Computer Modeling. 2004; 39: 209-217. https://doi.org/10.1016/S0895-7177(04)90008-8

[35] Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation. Machine Learning. 1999; 34: 177-210. https://doi.org/10.1023/A:1007506220214

[36] Pevzner L, Hearst M. A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics. 2002; 28(1): 19-36. https://doi.org/10.1162/089120102317341756

[37] Kazantseva A, Szpakowicz S. Topical Segmentation: a Study of Human Performance and a New Measure of Quality. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 211–220.

[38] Scaiano M, Inkpen D. Getting More from Segmentation Evaluation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 362–366.

[39] Appelt DE, Hobbs JR, Bear J, et al. Fastus: A finite-state processor for information extraction from real-world text. In: Proceedings of the IJCAI'93. 1993: 1172–1178.

[40] Marrero M, Sánchez-Cuadrado S, Morato Lara J, et al. Evaluation of Named Entity Extraction Systems. Advances in Computational Linguistics, Research in Computing Science. 2009; 41-47.

[41] Siefkes C. An Incrementally Trainable Statistical Approach to Information Extraction Based on Token Classification and Rich Context Models. Ph.D. Thesis. 2007.

[42] Singh S, Riedel S, Martin B, et al. Joint inference of entities, relations, and co-reference. In: Proceedings of the 2013 workshop on Automated knowledge base construction CIKM'13 22nd ACM International Conference on Information and Knowledge Management. 2013: 1-6.