Segmenting HTML pages using visual and semantic information

Georgios Petasis, Pavlina Fragkou, Aris Theodorakos, Vangelis Karkaletsis, Constantine D. Spyropoulos

Software and Knowledge Engineering Laboratory,

Institute of Informatics and Telecommunications,

National Centre for Scientific Research (N.C.S.R.) “Demokritos”,

P.O. BOX 60228, Aghia Paraskevi,

GR-153 10, Athens, Greece.

e-mail: {petasis, fragou, artheo, vangelis, costass}@iit.demokritos.gr

Abstract

The information explosion of the Web aggravates the problem of effective information retrieval. Even though linguistic approaches found in the literature perform linguistic annotation by creating metadata in the form of tokens, lemmas or part of speech tags, however, this process is insufficient. This is due to the fact that these linguistic metadata do not exploit the actual content of the page, leading to the need of performing semantic annotation based on a predefined semantic model. This paper proposes a new learning approach for performing automatic semantic annotation. This is the result of a two step procedure: the first step partitions a web page into blocks based on its visual layout, while the second, performs subsequent partitioning based on the examination of appearance of specific types of entities denoting the semantic category as well as the application of a number of simple heuristics. Preliminary experiments performed on a manually annotated corpus regarding athletics proved to be very promising.

1. Introduction

Nowadays, the vast amount of information available in the Web remains in a considerable degree an unexploited thesaurus of knowledge resources. This is due to the fact that, even though proposed methods for performing fast search, clever ranking, data analysis and web page indexing proved to be highly effective, the actual content of the web pages is very poorly exploited. In order to face this challenge, a number of approaches have been proposed, which try to attach metadata to web pages, annotating them with linguistic or even semantic information. Linguistic metadata typically involve the identification of tokens, part of speech tags, stems, lemmas and other syntactic information. However, these metadata are not enough if an in depth analysis of the content is required, in order to interpret and extract the facts and events described by the content. Effective content processing requires the presence of semantic metadata, which can be extremely useful for the efficient retrieval, extraction and interpretation of the information contained into a web page. In addition, it may help in overcoming problems such as search engine duplicate results elimination, efficient query refinement and high precision answering. Being a resource demanding and time consuming process, semantic annotation needs to be automated as much as possible through the automatic acquisition of semantic metadata according to a predefined semantic model.

An important subtask of semantic annotation is the semantic segmentation. Semantic segmentation can be defined as the task of identifying parts of a web page that refer to the same fact/event (topic), as defined by a semantic model. For example, web pages of a news portal may contain multiple news items on a single page, while pages from an electronic shop can contain multiple product descriptions per page. Semantic segmentation must disambiguate these multiple instances into areas describing a single item, by extracting and subsequently annotating the “semantic” structure of a web page with predefined categories (such as “news item” or “product”) according to a predefined semantic model. Various subtasks of information extraction can benefit from the results of semantic segmentation. Tasks like named entity recognition may benefit less, as they usually can be performed on a web page as a whole. However task like co-reference resolution and relation extraction can benefit more, as they can be applied in more confined areas that are semantically “coherent”, possibly limiting erroneous relations and thus increasing performance.

In this paper we present a novel method which performs automatic semantic segmentation on web pages. Most specifically, our method partitions the text on a web page into blocks, which refer to a single topic, and annotates identified blocks with predefined, domain specific, categories. The partitioning is based not only on the visual layout of a web page but also on information related to the distribution of named entity instances in the page. Starting from a detailed visual segmentation, the proposed method follows a two staged approach. During the first stage, a set of heuristics is applied to transform the visual segmentation into an initial semantic segmentation. This initial semantic segmentation is refined during the second stage by exploiting the distribution of named entities, with the help of a machine learning algorithm.

Preliminary results obtained from a manually annotated corpus about athletics, show that the segmentation performed by using information from named entities clearly outperform the segmentation that exploits only visual information, even if visual information is combined with heuristics, in order to increase its performance. The exploitation of the distribution of named entities, which is one of the main

innovative aspects of the presented approach, appears to significantly improve segmentation performance over heuristics that are frequently employed for the task. In addition, the elimination of the heuristics lead to a more adaptable system, as usually these heuristics are domain specific and quite often site specific. Finally, the importance of our work lies on the fact that it can be easily applied to a number of web-driven applications such as search engines, web-based question answering, web-based data mining as well as voice enabled web navigation.

The structure of the present paper is as follows. Section 2 provides an overview of related work found in the literature, Section 3 provides a description of the research project under which the aforementioned work took place as well as a detailed description of our proposed method, Section 4 provides experimental results while Section 5 provides conclusion results as well as future steps.

2. Related Work

Research approaches found in the literature use a variety of methods to accomplish the identification of the structure of a web page i.e. page segmentation in order to perform automatic semantic segmentation. A family of methods perform automatic extraction of segments i.e. blocks by defining appropriate patterns or grammars that contain the target data in web pages usually conforming to a common schema and repeating pattern mining or adopting pattern matching and string alignment techniques to refine extraction rules (Zhai and Liu, 2005; Zhai and Liu, 2006. Crescenzi et al, 2001; Crescenzi et al., 2002; Arasu and Garcia-Molina, 2003; Chang and Kuo, 2004; Chang et al., 2003). Such methods usually require (labeled or unlabelled) training examples. A second family of methods is based on machine learning requiring human labeling from each Web site that is interested in extracting data from (Kushmerick, 2000; Hsu and Dung, 1998; Muslea et al., 1999). Recent approaches (Wu et al.,2006; He et al., 2005; Song et al., 2004; Feng et al., 2005; Yang et al., 2006; Wang and Richard, 2007) also exploit the visual layout of a web page by looking in more depth in the actual structure of a web page. Such approaches are based on the observation that, blocks of interest usually appear in the central-main region of the web page and most specifically exhibit similar visual appearance. The exploitation of the structure of a web page is either performed in a "flat" manner i.e. by searching for repeated sequences of html tags or in a "hierarchical" manner by exploiting the DOM tree [1] or tag tree of the page. Blocks of interest are retrieved by exploiting the nodes of the tree and by applying heuristic rules, tree pattern matching, by examining the geometrical layout of the leaf nodes, linguistic features, spatial features (such as the position and size) and content features (such as the number of images and links) that are used to construct a feature vector for each block. The feature vector of each block is in his turn being used to train a model to assign importance to each block in the web page. Alternative approaches exploit the attributes of the DOM tree. Those approaches try to attribute a block importance value based on the existence of block features, spatial cues as well as on heuristic rules. They alternatively try to attribute a block importance value by defining content splitting delimiters that are applied in the DOM tree in order to create block nodes that are refined based on cue phrases appearing in the anchor text.

One among the most promising approaches is the one implemented by the VIPS algorithm (Cai et al., 2003; www.ews.uiuc.edu/~dengcai2/VIPS/VIPS.html) which tries to segment a web page by exploiting its visual layout after its rendering in a web browser. Using the visual information stored by the web browser into the DOM tree, VIPS identifies the segments that cover large areas of the rendered web page and then tries to find from the visual representation of the page horizontal or vertical lines that separate these segments. Separators denote the horizontal or vertical lines in a web page without visually crossing with blocks. The vision-based content structure of the page is constructed by recursively segmenting the already processed segments, while the granularity of segmentation can be controlled through a threshold that represents the coherence of the segment's content. A variation of the VIPS algorithm introduce an additional criterion i.e. the Degree of Isolation (Li et al., 2005) to describe the difference between a block and its other siblings and use language modeling to estimate a model for each block.

3. Method

The majority of existing algorithms works well in recognizing the visual layout of a web page. However, this proves to be insufficient for the semantic annotation, which requires semantic segmentation. Such produced metadata may be useful to create an effective representation of a web page to assist in tasks such as information retrieval, query refinement and information extraction. This is due to the fact that, for example, the information extraction's purpose is to identify segments of text containing semantic information, i.e. instances of named entities and relations. The identification of such instances and relations is more effective when it is performed into page segments describing a single event or "topic" instead of the whole page.

The method presented here goes a step further by making use of non-visual information for improving web page segmentation in order to be more compatible with semantic metadata information. More specifically we present a method consisting of two stages, which can act independently, in order to segment HTML web pages using visual and semantic information. As a first step, the VIPS visual segmentation algorithm is applied in order to visually segment each page as thorough as possible. With the help of heuristics applied on the results of visual segmentation, basic building blocks such as columns, paragraphs, headings, tables, headers and footers are identified.

As a subsequent step, the distribution of named entities in the text of a web page is exploited, in order to semantically segment the page. A named entity recognizer that can handle the domain of the web page is applied on the page. Then, with the use of machine learning, entities that can be attributed to refer to the same topic, are

[1] The Document Object Model (DOM) is an application programming interface (API) for valid HTML and well-formed XML documents. It defines the logical structure of documents and the way a document is accessed and manipulated.

grouped together. These groups form the basis for semantically segmenting the page, by classifying basic building blocks obtained during visually based segmentation that contain an entity group into the same semantic segment. Grouping is achieved by capturing the relations between named entity instances, as those are defined by a semantic model. Grouping is performed by the use of a machine learning algorithm which takes as input a web page and creates a representation containing special and proximity relations between named entity instances appearing in parts of texts. Such relations are not captured by single repetition of the same value if named entity instances. Relation instances of interest are identified by searching for repetition of specific types of named entity instances with respect to other ones. New topics may be also be identified by novel instances of predefined types of named entities. The result of the implementation of the machine learning algorithm on a web page is its segmentation into "semantic segments" according to the BOEMIE semantic model.

3.1 Visual segmentation with the use of heuristics

The first stage of our method, which is based on heuristics operating on the results of visual segmentation, tries to identify a web page's content structure by exploiting spatial attributes such as relevant position or size. The VIPS algorithm is applied, in order to perform a thorough segmentation based solely on visual information. The segmentation detail is controlled through the granularity parameter of VIPS, which controls the coherence of identified segments based on visual perception. By using the maximum granularity value, this process aims to identify the smallest possible visual segments, which will be concatenated through simple heuristics to form the basic building blocks, such as headers, footers, columns, headings, paragraphs, tables, images and captions. VIPS uses placement and size information, placed by a web browser in the DOM tree nodes that correspond to visual structures, in order to annotate the DOM tree with segment information. From the resulting DOM tree, our method tries to identify the "main" or "central" region of a page, usually denoted by the wider column whose length occupies a significant percentage of the total page length. In the majority of cases this column is represented by the <td> HTML tag, but in cases where this assumption fails, the DOM tree node having the greatest text length (in characters) is used instead. After detecting the main region, the leaves of the sub-tree of the main region node that contain at least some text are examined in order to detect titles and paragraphs. The assumption is that such nodes have maximum coherence since they were retrieved at the final level of VIPS segmentation. Leaves which correspond to image captions or contain linked text only are omitted in order to result in a set of nodes appearing in the main region and contain only text i.e. nodes that are either titles or paragraphs. In order to identify such cases the following heuristics were applied: (a) if the text of the node is up to 5 words, it is considered as a title (b) if it contains more than 20 words it is considered as a paragraph (c) if it contains a number of words in the range from 6 to 20 words, it is considered as a paragraph in case it ends with a dot, otherwise it is considered as a title. The procedure finishes after all possible nodes have been parsed, while the output consists of the identified titles and paragraphs.

The choice of the aforementioned heuristics is based on results obtained from the observation of web pages contained in the corpus at our disposal regarding athletics, as well as the statistical calculation of the number of words appearing in titles and paragraphs in each of those pages.

3.2 Transforming visual segments into semantic segments

While the first stage of our method focuses on the exploitation of spatial attributes, the second one is focused on identifying segments which describe the same topic, with respect to the thematic domain of a web page. For example, in pages containing news about athletic events, the semantic model of the domain may dictate the recognition of segments describing a single sport competition, in a page describing all sport competitions that took place in a specific event date. Topic segmentation is accomplished by parsing titles and paragraphs in order to identify changes in the frequency of appearance and distribution of specific types of named entities, adopting a machine learning approach. The assumption here is that the presence of specific types of entity instances as well as the repetition of some of them may prove to be determinative for the presence of the same topic. Consequently, the change of the instance values of specific types of entities may signal change of topic. Spatial and proximity information of named entity instances play a major role to the signaling of the presence or change of the same topic.

In order to perform segmentation, a semi-supervised machine learning approach is followed, which, given an unknown web page, it tries to identify portions of texts that refer to a single topic. This identification is based on an already learned model. To construct such as model, a number of manually annotated web pages with the desired semantic segments are used to form a training corpus. Each page belonging to the training corpus is processed in order to form one or more feature vectors each of which represents a different topic. A feature vector is constructed for every title or paragraph containing those features that may signal the existence or not of the same topic. The feature vector contains information regarding the type of the node and the frequency of occurrence of named entities in the node. While all identified types of named entities can be used in the vector, a reduction of the used types is also possible, if some types can be identified as being more important than others in advance. For example in the athletics domain, if topic is assumed to be news about sport competitions, entities like sport names, athlete names and gender of athletes can be more significant for the task, than entities like locations or nationalities. Each feature vector also contains contextual information, by containing the equivalent information found into previous node(s). The feature vector finally contains the number of common named entity instance values with those of the previous node(s).

3.3 Case study: the BOEMIE R&D project

The proposed semantic segmentation method has been applied in the context of the BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction) R&D project (http://www.boemie.org).

BOEMIE implements ontology-based information extraction systems which extract metadata information compliant to an ontology from multimedia content regarding athletic events. Driven by domain-specific multimedia ontologies, the BOEMIE information extraction system is able to identify high-level semantic features in images, videos, audios and texts, and then fuse modality specific extracted information to interpret multimodal resources.

When dealing with web pages about news, usually obtained from web sites of official associations like IAAF, topic detection mainly concerns the detection of news items describing a single sport. A news item is usually represented by an (optional) title, followed by one or more paragraphs that refer to a specific sport, taking place during either a specific event or during multiple events. News-item detection is performed by parsing the titles and paragraphs in order to identify appropriate units. Units are usually portions of a web page describing the athlete's participating in a specific sport during a specific event or the athlete's participating in more than one event when the sport in question remains the same.

The named entities defined in the semantic model can be grouped as follows: (a) athlete_name, age, gender, nationality (corresponding to the information regarding an athlete), (b) performance and ranking (of an athlete in a specific round of a specific sport), (c) round_name and heat_name (corresponding to a specific round with or without the information of the heat of a specific sport) (d) sport_name, (corresponding to a specific sport), (e) event_name, country (corresponding to a specific event and country in which the event takes place) (f) city, date (corresponding to a specific city or date in which a specific sport or event takes place). Our assumption is that, when during the text, a change of an instance of a named entity such as the sport name, the athlete name and/or the gender takes place, this change may signal the appearance of a new news item. Figure 1 depicts a part of the domain ontology for the text modality in BOEMIE.

3.3.1 Extracting semantic segments

In order to adapt the proposed method to the domain of athletics, we have decided to limit the entity types involved to occurrences of named entities with types sport name, athlete name and gender. The selection of those types was based on the following observations: (a) the most frequent information found in paragraphs and titles belonging to a news item is the athlete and sport named entities, (b) paragraphs constituting a news item usually describe the progress of a competition, thus contain a number of common sport and athlete name instances, (c) the gender information signals the description of men or women competitions, which are considered as different sports and thus news items (topics), (d) the information of performance can only act indirectly to the identification of a news item (i.e. through the range of observed values), (e) change on the sport name usually corresponds to a different sport competition.

It is worth mentioning that our feature vector takes advantage of the actual names of the named entities of interest. Thus, it focuses on the change of values of specific types of named entities but not on the change of the types on named entities used.

From the portion of a webpage which is depicted in figure 2, the following feature vectors are produced from examining its nodes

- **Title:** Title , athlete_instance_1 = "Richards", #athlete_instances = 1
- **Paragraph 1:** Paragraph, sport_instance_1 = "400m", athlete_instance_2 = "Sanya Richards", gender_instance_1 = "women", #athlete_instances = 1, #sport_instances = 1, #gender_instances = 1, *common_information_with_title* = (athlete_instance_1, #=1)
- **Paragraph 2:** Paragraph, athlete_instance_3 = "Richards", #athlete_instances = 1, *common_information_with_title* = (athlete_instance_1, #=1), *common_information_with_previous_paragraphs* = (athlete_instance_2, #= 1)
- **Paragraph 3:** Paragraph, athlete_instance_4 = "Richards", #athlete_instances = 1, *common_information_with_title* = (athlete_instance_1, #=1), *common_information_with_previous_paragraphs* = (athlete_instance_2, athlete_instance_3, #= 2)
- **Paragraph 4:** Paragraph, athlete_instance_5 = "Lashawn Merritt", athlete_instance_6 = "Angelo Taylor", sport_instance_2 = "400m"', sport_instance_3 = "400 Hurdles", gender_instance_2 = "men, #athlete_instances = 2, #sport_instances = 2, #gender_instances = 1, *common_information_with_title* = (nothing),*common_information_with_previous_paragraphs* = (sport_instance_1, #= 1)

In order to learn a model for semantically segmenting a web page, the CRF++ algorithm (http://crfpp.sourceforge.net/) was chosen, due the fact that it has been proved to be a very effective framework for building probabilistic models to segment and label sequential data (Sha and Pereira, 2003). CRF++ is a simple, customizable, and open source implementation of Conditional Random Fields (CRFs) for segmenting/labeling sequential data. CRF++ has been designed for generic use, and has been applied on a variety of NLP tasks, such as Named Entity Recognition, Relation Extraction and Text Chunking.

4. Experiments

The experiments of our algorithm were performed on a manually annotated corpus. This is due to the fact that a suitable benchmark corpus for this specific scientific area is not publicly available. The collection of the corpus, as well as the way documents were annotated are described in the following subsections.

4.1 The dataset

In order to evaluate our method we collected a corpus which contains 100 web pages taken from eight different web sites, which are the following:

- IAAF (http://www.iaaf.org) from the www.iaaf.org site, pages containing news regarding athletic events were collected, where both text and images may appear within it. (37 web pages)

- USA Track & Field (http://www.usatf.org/), pages containing news regarding athletic events, where both text and image may appear within it were collected. (28 web pages)
- news.bbc.co.uk/, (12 web pages)
- http://www.sportinglife.com/, (10 web pages)
- http://www.scc-events.com/ (4 web pages)
- http://sportsofworld.com/. (9 web pages)

In all web pages manually annotation of the boundaries of the blocks of interest i.e. the "news items" was performed using an annotation facility provided by the Ellogon platform (http://www.ellogon.org). More specifically, the annotation tool of Ellogon was appropriately configured in order to annotate the following entities:

- Semantic categories: news item, sport
- Columns: column left, column main, column right
- Tables: table, table column, table row
- Headings: caption, title
- Page areas: footer, header, paragraph, sentence, other region
- Semantic model entities: athlete name, age, gender, nationality, performance, ranking, round name, heat name, city, date, country, sport name, stadium name and event name which correspond to the BOEMIE semantic model. It is worth mentioning that overlapping annotation may exist within a web page. A screenshot of the annotation tool is given in Figure 3.

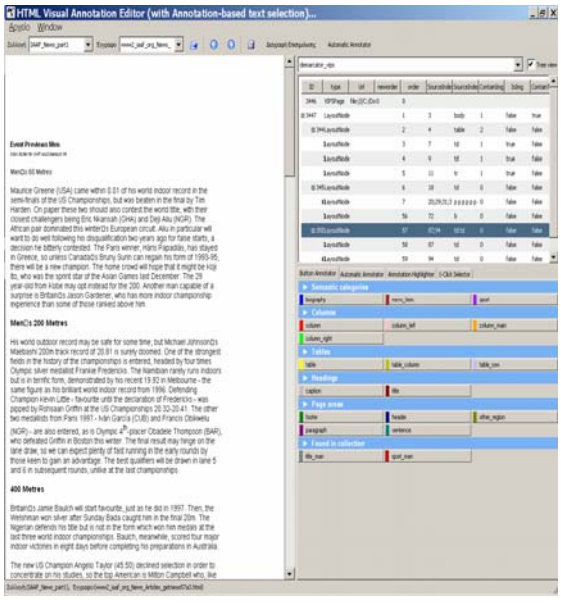

Figure 3: Annotation of a single web page

4.3 Evaluation

In order to evaluate the proposed method, we measured its performance on the task of news item identification, in terms of the Precision, Recall and F-measure metrics. More specifically, precision is defined as "the number of the estimated news items which are actual news items" divided by "the number of the estimated news items, returned by the method". Recall is defined as "the number of the estimated news items which are actual news items" divided by "the number of the true news items". F-measure was defined as the double of the product of precision and recall divided by their corresponding sum. More specifically, we examine the correct assignment of boundaries of our algorithm with those appearing in the gold corpus with respect to the paragraph's sequence of appearance within the page.

Two different approaches were examined in order to evaluate the correct identification of news items. The first approach applies a simple heuristic on top of the heuristics described in Section 3.1. Making the assumption that news items usually consist of a title followed by one or more paragraphs, the heuristic approach uses the identified information about paragraphs and titles, to form news items. Each identified title is merged with all paragraphs until the next title (or the end of the page) to form a news item. Thus, changes in news items are signaled only by the presence of a new title.

The second approach consists of the application of the machine learning algorithm described in Section 3.2, on top of the heuristics described in Section 3.1. A five fold cross validation was performed on the machine learning approach. Both the heuristics as well as the combined approach (heuristics and machine learning algorithm) were applied on the web pages DOM trees produced by VIPS, in order to evaluate the correct identification of news items. The results obtained by both evaluations are listed in the table below.

	Only heuristics	**Heuristics & machine learning algorithm**
Precision	40.83%	71.19%
Recall	18.64%	70.87%
F-measure	25.59%	71.01%

Table 1: Evaluation results in news item detection with heuristics and combination of heuristics and machine learning

From the obtained results we can see that the mere application of heuristics is not enough for the detection of news items. This can be attributed to several facts. The most important one is that the heuristics chosen proved to be too general (not domain specific because they do not exploit the presence of named entity instances that appear within them which proves to be the "key" for detecting news items) and able to capture a small portion of cases, the majority of those belonging to the IAAF site. A second important factor is the assumption that a news item consists of a title and a number of paragraphs below it proves to be true in few cases. This is due to the fact that, the paragraphs appearing below a title practically prove to belong to more than one news items. The low performance of the heuristics can also be attributed to the variety of web pages layout. Web pages layout proves to strongly differ from site to site but also from page to page within the same site.

The use of semantic features via machine learning proves to significantly improve news item detection. This can be attributed to several factors: (a) the high efficiency of the CRF algorithm in segmenting and labeling sequential data (b) the choice of the types of the named entities chosen to form the feature vector. This is due to the fact that the sport name entity in web pages containing news is a strong indicator of the appearance of a novel

news item, although it tends to appear only once inside a news item, and not necessarily in the first paragraph of a news item. The same also holds for the appearance of novel athlete name instances which indicate that the subject in question refers to a different sport or event. Finally the change of gender name instance also signals the change of topic, which is indirectly assisted by the change of athlete name instances.

It is worth mentioning that, news item detection in web pages containing news must be dealt as a difficult problem. This can be attributed to several factors. The first and probably most important reason is the fact that those pages have quite different layouts. More specifically, pages belonging to this category may either contain only plain text i.e. paragraphs without section titles, or containing sections accompanied with section titles. In the first case the news item detection is defined as the problem of finding the appropriate number of paragraphs that corresponds to a news item. In the second case, the text portion appearing between two section titles may actually belong to more than one news items due to the fact that the description of more than one sport for the same event, or the description in more than one event for the same sport may take place. The aforementioned situation is aggravated by the fact that it is difficult even for a domain expert to decide upon a single topic for such cases. For example, there are documents regarding athletics where in their beginning, the "highlights" of more than one sport (along with their top-scoring athletes) are described in the same paragraph, making impossible to classify this paragraph with a single topic.

5. Conclusions – Future Work

In the current paper, we presented a novel approach to the problem of automatic semantic segmentation, which tries to automatically identify portions of text within a web page corresponding to semantic categories, in order to create semantic annotation to web pages. The produced semantic annotation corresponds to semantic segments, each of which captures a single topic i.e. "news" item, where named entity instances related to the topic are contained. This type of segmentation provides an overview of the content of a web page, with respect to the semantic model in use. The innovation of our method lies in the fact that it takes advantage of the semantic information, as expressed by the named entity instances, and combines it with visual layout information to perform semantic segmentation.

The examination of the results produced, especially concerning the second stage of our method, lead to the conclusion that the use of specific types of names such as sport and athlete names for web pages describing athletic events proved to be beneficial. A significant property of our method is that it can be applied to additional topic areas, possibly significantly different from athletics, as far as a semantic model is appropriately defined.

As future work, we plan to apply our algorithm to a larger corpus of web pages, exhibiting even greater layout variance, possibly originating from a larger variety of web sites, by exploiting better the large corpus collected in the context of the BOEMIE project. In order to accomplish this as well as to higher scalability regarding the domain in use, we plan to combine our method with others used for topic change i.e. text segmentation such as those of (Kehagias, Fragkou and Petridis, 2004) and (Utiyama and Isahara, 2001). Text segmentation methods benefit from the fact that they are domain independent and they are based on the statistical distribution of words within paragraphs or sentences. Those methods can be used as a preprocessing step. Alternatively, they can ground the calculation of their statistical distributions on specific types of words i.e. named entities and more specifically on the degree of change of the types of named entities and/or the degree of change of the actual names.

Regarding heuristics, we plan to make use of the information of the named entity instances contained in titles and paragraphs i.e. parts of texts to which they apply. We intend to examine the use of new features, such as different types of named entities that can further improve the performance of our algorithm. We also plan to exploit information resulting from the application of shallow parsing and/or co reference resolution in order to identify portions of text referring to the same topic as well as to exclude "false alarm" instances. Furthermore, we additionally consider examining alternative learning algorithms, besides conditionally random fields. Finally, due to the fact that, the measures of Precision, Recall and F-measure penalize equally every inaccurately estimated news item boundary whether it is near or far from a true segment boundary, we plan to evaluate our method using the Beeferman's P_k metric (Beeferman, 1999) as well as the WindowDiff metric (Pevzer & Hearst, 2002). Those measures are successfully used for the evaluation of change of topic and text segmentation algorithms.

6. References

Arasu, A., Garcia-Molina, H. (2003). Extracting structured data from Web pages. In the *International Conference on Management of Data Proceedings, session: Data integration and sharing II*, pp: 337—348.

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

Cai, D., Yu, S., Wen, J-R., Ma, W-Y. (2003). Extracting Content Structure for Web Page based on Visual Representation". In the *Fifth Asia Pacific Web Conference*.

Cai, D., Yu, S., Wen, J-R., Ma, W-Y. (2003). VIPS: a Vision-based Page Segmentation Algorithm. Microsoft Technical Report (MSR-TR-2003-79).

Chang, C-H., Kuo, S-C. (2004). OLERA: Semisupervised Web-Data Extraction with Visual Support. In *IEEE Intelligent Systems*, 19(6), pp. 56--64.

Chang, C-H., Hsu, C-N., Lui, S-C. (2003). Automatic information extraction from semi-structured Web pages by pattern discovery. *Web retrieval and mining*, 35(1), pp. 129 -- 147.

Crescenzi, V., Mecca, G., Merialdo, P. (2001). RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In *Proceedings of the 27th International Conference on Very Large Data Bases*, pp. 109 -- 118.

Crescenzi, V., Mecca, G., Merialdo, P. (2002). The RoadRunner Project: Towards Automatic Extraction of Web Data. In *Proceedings of the International*

Workshop on Adaptive Text Extraction and Mining in conjunction with the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001) Symposium on Applied Computing, pp. 1108 -- 1112.

Feng, J., Haffner, P., Gilbert, M. (2005). A learning approach to discovering Web page semantic structures. In the *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, pp. 1055--1059.

He, Z., Gao, Z., Xu, H., Qu, Y. (2005). DeSeA: A Page Segmentation based Algorithm for Information Extraction. In the *Proceedings of the First International Conference on Semantics, Knowledge and Grid*, pp. 14--14.

Hsu, C-N., Dung, M-T. (1998). Generating Finite-State Transducers for Semi-structured Data Extraction from the Web. In *Information Systems*, 23(9), pp. 521--538.

Kehagias, Ath., Fragkou P. and Petridis V. (2004). A Dynamic Programming Algorithm for Linear Text Segmentation. *Journal of Int. Information Systems*, 23, pp. 179--197.

Kushmerick, N. (2000). Wrapper Induction: Efficiency and Expressiveness. In *Artificial Intelligence*, nos. 1-2, pp. 15--68.

Li, S., Huang, S., Xue, G-R., Yu, Y. (2005). Block-based Language Modeling Approach towards Web Search. In *Proceedings of the seventh Asia-Pacific Web Conference*, pp. 170--182.

Muslea, I., Minton, S., Knoblock, C. (1999). A Hierarchical Approach to Wrapper Induction. In *Proceedings of the Third Annual Conference on Autonomous Agents*, pp. 190--197.

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

Sha F., and Pereira F. (2003). Shallow Parsing with Conditional Random Fields. In *Proceedings of HLT-NAACL 2003*, Edmonton, Canada, 2003, pp. 213-220.

Song, R., Liu, H., Wen, J-R., Ma, W-Y. (2004). Learning important models for web page blocks based on layout and content analysis. In *ACM SIGKDD Explorations Newsletter*, 6(2), pp.14 --23.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain - independent text segmentation. *In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 491--498.

Wang, Y., Richard, R. (2007). Rule-based Automatic Criteria Detection for Assessing Quality of Online Health Information. In the *International Conference Addressing Information Technology and Communications in Health (ITCH)*.

Wu, C., Zeng, G., Xu, G. (2006). A Web page Segmentation Algorithm for extracting product information. In the *IEEE International Conference on Publication*, pp. 1374--1379.

Yang, X., Xiang, P., Shui, Y. (2006). Semantic HTML Page Segmentation using Type Analysis. In the *1st International Symposium on Pervasive Computing and Applications*, pp. 669--674.

Zhai, Y., Liu, B. (2006). Structured Data Extraction from the Web Based on Partial Tree Alignment. In the *IEEE Transactions on Knowledge and Database Engineering*, 18(12), pp. 1614--1628.

Zhai, Y., Liu, B. (2005). Web data extraction based on partial tree alignment**.** In *Proceedings of the 14th international conference on World Wide Web*, pp: 76--85.

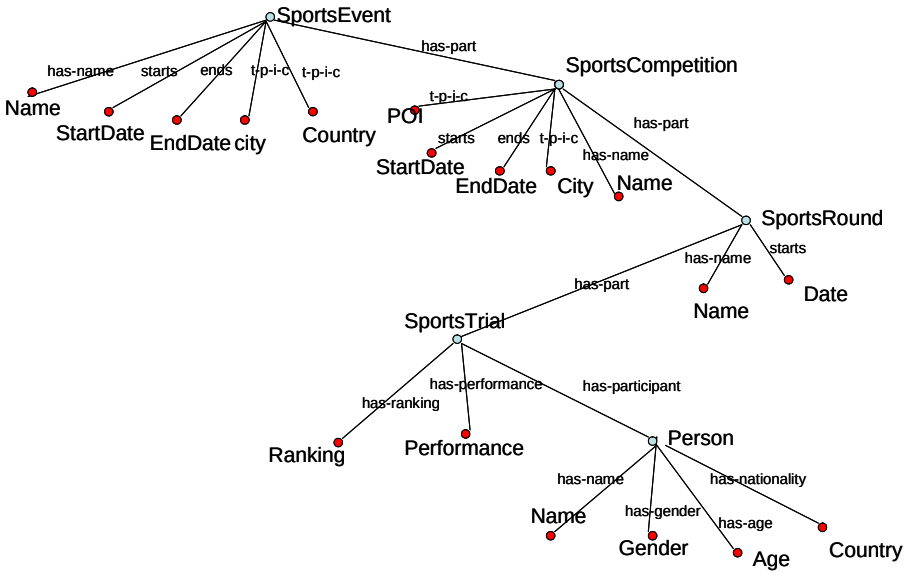

Figure 1: Part of the BOEMIE domain ontology

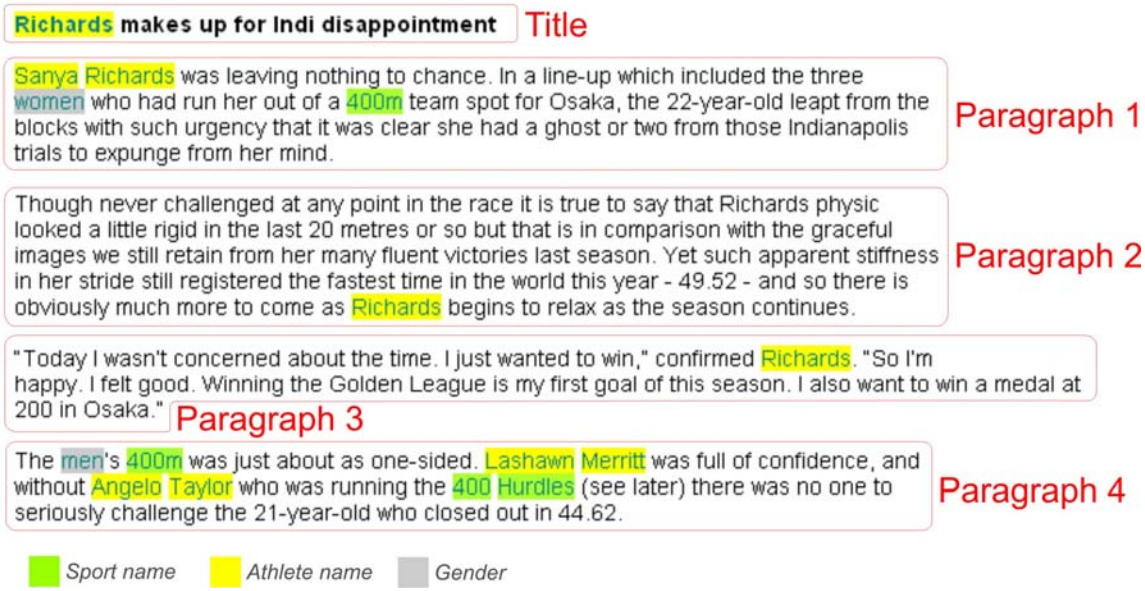

Figure 2. Example of an annotated web page