Semantic Annotation using Horizontal and Vertical Contexts

Mingcai Hong, Jie Tang, and Juanzi Li

Department of Computer Science & Technology, Tsinghua University, 100084. China.
{hmc, tj, ljz}@keg.cs.tsinghua.edu.cn

Abstract. This paper addresses the issue of semantic annotation using horizontal and vertical contexts. Semantic annotation is a task of annotating web pages with ontological information. Information on a web page is usually two-dimensionally laid out, previous semantic annotation methods that view a web page as an 'object' sequence has limitations. In this paper, to better incorporate the two-dimensional contexts, semantic annotation is formalized as a problem of block detection and text annotation. Block detection is aimed at detecting the text block by making use of context in one dimension and text annotation is aimed at detecting the 'targeted instance' in the identified blocks using the other dimensional context. A two-stage method for semantic annotation using machine learning has been proposed. Experimental results indicate that the proposed method can significantly outperform the baseline method as well as the sequence-based method for semantic annotation.

1. Introduction

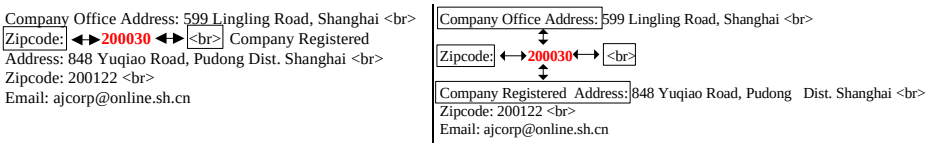
Semantic web requires annotating existing web content according to particular ontologies, which define the meaning of the words or concepts in the content [1]. In recent years, semantic annotation has received much attention in the research community. Many methods have been proposed, for example, manual annotation, rule learning based annotation, and machine learning based annotation.

Conversional automatic annotation methods typically convert the web page into an 'object' sequence and utilize information extraction (IE) techniques to identify a sub-sequence that we want to annotate (i.e. targeted instance). (Here, the object can be either natural language units like token and text line, or structured units indicated by HTML tags like "<table>" and "<image>"). However, information on a web page is usually two-dimensionally laid-out and should not be simply described as a sequence. Figure 1 shows an example of document.

In this example, the targeted instance is the highlighted text "200030". In terms of the sequence-based method, the snippet can be viewed as a token sequence and the task is to identify the sub token sequence "200030" (cf. Figure 2 (a), where "
" indicates a line break). In the identification, a usual approach will identify the start position and the end positions based on the context prior to and next to the targeted instance, e.g. "Zipcode:" and "
". Unfortunately, in the example, the method will confuse the text "200122" with "200030" because they have the same context.

...
4. Company Office Address: 599 Lingling Road, Shanghai
Zipcode: 200030
Company Registered Address: 848 Yuqiao Road, Pudong Dist. Shanghai
Zipcode: 200122
Email: ajcorp@online.sh.cn
...

Fig 1. Example of document



(a) One-dimensional context (b) Two-dimensional context

Fig 2. One-dimensional context vs. Two-dimensional context

An alternative method is to take into consideration of both the horizontal context and the vertical context (cf. Figure 2 (b)). For the targeted instance “200030”, its vertical contexts (including above context “Company Office Address:” and below context “Company Registered Address:”) can be used to distinguish it from instance “200012” and its horizontal contexts (including left context “Zipcode:” and right context “
”) can be used to identify its start position and end position.

In this paper, to better incorporate the horizontal context and the vertical context, a two-stage method for semantic annotation is proposed in this paper. We formalize semantic annotation as that of block detection and text annotation. We propose to conduct semantic annotation in the two-stage fashion. We view the tasks as classification and propose a unified statistical learning approach to the tasks, based on Support Vector Machines (SVMs). The proposed method has been applied to a commercial project TIPSI, which is aimed at annotating the company annual reports from Stock Exchange. We used company annual reports from Shanghai Stock Exchange for experimentation. Our experimental results indicate that the proposed two-stage methods perform significantly better than the baseline methods for semantic annotation. We observed +11.4% and +16.3% improvements (in terms of F1-measure) than the rule-based method and sequence-based method.

The rest of the paper is organized as follows. In section 2, we introduce related work. In section 3, we describe our approach to semantic annotation using horizontal and vertical contexts. In section 4, we use the annotation of company annual reports as a case study to explain one possible implementation. Section 5 gives our experimental results. We make concluding remarks in section 6.

2. Related Work

Related work can be summarized into three categories: annotation using rule induction, annotation as classification, and annotation as sequential labeling.

Many existing semantic annotation systems make use of rule induction to automate the annotation process (also called as ‘Wrapper’ induction, see [2]). For example,

Ciravegna et al propose a rule learning algorithm, called LP^2, and have implemented an automatic annotation module: Amilcare [3]. The module can learn annotation rules from training data. The learned rules can then be used to annotate un-annotated documents. Amilcare has been used in several annotation systems, for instance, S-CREAM [4]. See also [5].

Another method views semantic annotation as classification, and automates the processing by employing statistical learning approaches (e.g. Support Vector Machines (SVMs) [6]). It defines features for each candidate instance and tries to learn a classifier that can detects the targeted instances from the candidate ones.

Different from the rule induction and the classification based methods, sequential labeling enables describing the dependencies between targeted instances in the semantic annotation. The dependencies can be utilized to improve the accuracy of the annotation. For instance, [7] proposes utilizing HMM in semantic annotation.

Much of the previous work converts the web page into an 'object' sequence (e.g. token sequence or text-line sequence) and utilizes information extraction (IE) techniques for identifying the targeted instance.

3. A Two-stage Approach using Horizontal and Vertical Contexts

In this paper, by *context*, we mean the surrounding information of the targeted instance. By *horizontal context*, we mean information left to and right to the targeted instance (e.g., the previous tokens and the next tokens). And by *vertical context*, we mean information above and below of the targeted instance (e.g., the previous lines and the next lines). For semantic annotation, we target at detecting the instances from a document and annotating each of the instances by a concept in a particular ontology.

We adopt a strategy of divided-and-conquer and formalize the problem of two-dimensional contexts based semantic annotation as that of block detection and text annotation. A *block* is a specific informative unit in a document. It can be defined by different granularity, e.g. text line, section, or paragraph. We also assign a label to each block. The assigned label corresponds to a concept in the ontology, implying that the block contain at least one instance of the concept. A block can have multiple labels indicating the block contains instances of different concepts. A block can also have no label (i.e. "*none*") indicating that it contains no instance of any concept. The block can be laid horizontally or vertically. For facilitating the later explanation, we use vertically laid block as example hereafter.

In our two-stage approach, for block detection, a document is first viewed as a block sequence. For each block, we make use of its vertical context to detect its label. For text annotation, we view each identified block as an 'object' sequence and employ the horizontal context to detect the targeted instance.

In this work, we try to propose a general approach for semantic annotation. As case study, we work on annotating company annual reports. We only handle the annual reports in plain text format, i.e. non-structured data. We define a block as a text line, because in our experiments, statistic shows that 99.6% of the targeted instances are in one single text line (the statistic was conducted on the 3,726 experimental reports).

We formalize the two detection tasks as classification and employ a supervised machine learning approach. In block detection, we detect the label of each block using one classification model (the label corresponds to a concept in the ontology). In text annotation, we identify the start position and the end position of an instance using two classification models, respectively.

4. Annotating Company Annual Report using Two-Stage Approach

To evaluate the effectiveness of the proposed approach, we applied it to a practical project TIPSI. In TIPSI, we are aimed at annotating the company annual reports from Shanghai Stock Exchange (SSE).

A company annual report generally consists of fourteen sections, including "Introduction to Company", "Company Financial Report", etc. A comprehensive annotation for the company annual reports should annotate company basic information, financial information, and directorate information, etc. Due to space limitation, we will only describe the annotation of the first part (i.e. Section "Introduction to Company") and omit details of the rest. Section "Introduction to Company" contains company information such as *Company-Chinese-Name*, *Legal-Representative*, *Company-Secretary*, and *Office-Address*. (See Section 5 for details.)

We make use of Support Vector Machines (SVM) as the classification model [6]. SVM-light, which is available at http://svmlight.joachims.org/, is employed in our experiments. We choose linear SVM in both block detection and text annotation tasks. We use the default values for the parameters in SVM-light.

In the rest of the section, we will explain processes of block detection and text annotation and feature definition in the two processes.

4.1 Block Detection

Detections of different types of blocks are similar problems. We view block detection as classification. For each concept, we train a SVM model to detect whether a block contains instance(s) of that concept. A text line is viewed as a block in this task. The key issue then is how to define features for effectively learning and detecting. In all detection models, we define features at token level and line level. In the next sub-section, we will take *ccn* as example to explain the feature definition in block detection models. Features used in *ccn* block detection model are:

Positive Word Features: The features represent whether or not the current line contains words like "公司" and "中文". The words are usually used in the *ccn* block.

Negative Word Features: The features represent whether or not the current line contains words like "英文", "电话". These words are usually used in the other types of blocks and should not be included in the *ccn* block.

Special Pattern Features: A set of regular patterns is defined to recognize special patterns, such as email address, telephone number, fax number, URL. Each of the features respectively represents whether or not the current line contains one type of the special patterns.

Line Position Feature: The feature represents the line number of the current line. *ccn* block is usually placed in the first lines.

Number of Words Feature: The feature stands for the number of words in the current line.

The features above are also defined similarly for the previous line and the next line.

4.2 Text Annotation

An identified block contains at least one instance. We then try to identify the start position and the end position of the targeted instance. We view the problem as that of 'reverse information extraction' and employ two SVM models to perform the task. We also use the annotation of *ccn*'s instance as example in our explanation. Features used in *ccn* text annotation model are:

Token Features: The features respectively represent the specific tokens in the previous four positions, the current position, and in the next two positions. We define features using four previous tokens and only two next tokens. This is because our preliminary experiments show that the previous tokens seem more important in our annotation tasks.

Special Pattern Features: The features represent whether or not the current token contains a special pattern such as email address, telephone number, fax number, URL.

5. Experimental Results

5.1 Experiment Setup

We collected company annual reports from Shanghai Stock Exchange (http://www.sse.com.cn). We randomly chose in total 3,726 annual reports from 1999 to 2004. To evaluate the effectiveness of our approach, we extracted the Section "Introduction to Company" from each annual report for experiments.

In all the experiments, we conducted evaluations in terms of precision, recall and F1-measure. For block detection, we conduct evaluation at the line level. For the text annotation tasks, we perform evaluation at the 'instance' level.

We use the rule based annotation as baseline. The rules were defined according to the most useful features in the SVM models. For example, the rule to annotate *ccn* is "Token sequence starts after '*company Chinese name:*' and ends with '*Co., Ltd.*'".

We also compare the proposed approach with the sequence-based method. In this method, an annual report is viewed as a token sequence, and two SVM models are used to detect the start position and the end position, respectively. The same feature sets are used as that in the proposed approach for text annotation.

5.2 Experimental Results

We randomly split the data set into two 50:50 subsets, one for training and the other for test. We then conducted the experiment in the following way. First, we used the SVM models to detect the type of each block (i.e. text line) and assign (a) label(s). Next, based on the output of block detection, we used two SVM models to detect and annotate the target instances. Block predicted as "*none*" were skipped. For each experiment, we repeated the split and conducted the experiments for ten times. We used the average results as the experimental result. We also made comparisons with the baseline methods described above.

Table 1 shows the experimental results on the data set. Baseline and Sequence denote the baseline method and the sequence-based method defined above, respectively. Our Approach denotes the proposed approach. Pre., Rec., and F1 respectively represent the precision, recall, and F1-measure.

Table 1. Performance of annual reports annotation (%)

Annotation Task		**Pre.**	**Rec.**	**F1**	**Annotation Task**		**Pre.**	**Rec.**	**F1**
Company Chinese Name (ccn)	Baseline	97.4	86.8	91.8	Registered Address (caddr)	Baseline	91.6	83.3	87.3
	Sequence	97.6	87.4	92.2		Sequence	86.3	63.9	73.6
	Our Approach	97.4	90.1	93.6		Our Approach	88.3	92.0	90.1
Company English Name (cen)	Baseline	74.1	70.1	72.0	Office Address (coffice)	Baseline	88.6	88.7	88.6
	Sequence	92.5	87.8	90.1		Sequence	83.6	64.0	72.5
	Our Approach	94.8	91.1	92.9		Our Approach	89.2	90.2	89.7
English Name Abbreviation (ceabbr)	Baseline	95.4	78.8	86.3	Zip of Office Address (czip)	Baseline	88.6	78.9	83.5
	Sequence	97.9	85.9	91.5		Sequence	73.7	93.9	82.5
	Our Approach	92.7	90.7	91.7		Our Approach	96.7	93.4	95.0
Legal Representative (delegate)	Baseline	93.4	92.2	92.8	Website (curl)	Baseline	91.2	69.1	78.6
	Sequence	96.0	94.7	95.4		Sequence	61.7	89.1	72.9
	Our Approach	95.8	96.8	96.3		Our Approach	90.3	93.0	91.7
Company Secretary (sperson)	Baseline	89.3	88.9	89.1	Email of Company (cemail)	Baseline	94.1	45.8	61.6
	Sequence	94.9	88.4	91.5		Sequence	89.6	34.7	50.1
	Our Approach	87.9	94.0	90.8		Our Approach	93.1	87.1	90.0
Tel. of Secretary (stel)	Baseline	88.8	75.4	81.6	Newspaper (newspaper)	Baseline	88.5	70.4	78.4
	Sequence	51.1	82.7	63.2		Sequence	97.8	95.2	96.5
	Our Approach	91.5	96.1	93.7		Our Approach	97.6	98.1	97.8
Fax (sfax)	Baseline	92.3	91.2	91.7	Stock Name (sname)	Baseline	88.3	77.0	82.3
	Sequence	55.5	83.9	66.8		Sequence	94.8	86.1	90.2
	Our Approach	96.3	96.5	96.4		Our Approach	91.2	95.3	93.1
Address of Secretary (saddr)	Baseline	92.2	91.3	91.7	Stock Code (sno)	Baseline	96.2	86.3	91.0
	Sequence	58.4	73.6	65.1		Sequence	94.5	90.3	92.3
	Our pproach	95.8	97.0	96.4		Our Approach	95.5	95.2	95.3
Email of Secretary (semail)	Baseline	75.1	81.0	77.9	Average	Baseline	89.7	79.7	83.9
	Sequence	41.4	66.2	50.9		Sequence	80.4	80.5	80.4
	Our Approach	93.8	95.2	94.4		Our Approach	**93.4**	**93.6**	**93.5**

We see that our method can achieve good performances in all the tasks. For each annotation task, our approach significantly outperforms the baselines as well as the sequence-based methods. Now, we make discussion for the experimental results.

(1) Improvements over baseline method. The baseline method suffers from low recall in most of the annotation tasks, e.g. *cemail*, *curl*, and *newspaper*, although its precision is high. This is due to a low coverage of the rules. Our approach

outperforms the baseline method by 11.4% in terms of F1-measure. This also indicates that the features used in block detection and text annotation are effective.

(2) Two-dimensional context vs. One-dimensional context. In annotation of *ccn*, *cen*, *ceabbr*, *delegate*, *sperson*, *newspaper*, *sname*, and *sno*, the sequence-based method achieved high performance. This is because these fields are distinguishable by using only the horizontal context. While in the other annotation tasks, the sequence-based method suffers from lack of context and results in poor performance, even poorer than the baseline. It confirms us that accurate semantic annotation on company annual reports requires not only horizontal context, but also vertical context. Our approach benefits from the usage of both horizontal and vertical contexts.

(3) Error analysis. We conducted error analysis on the results of our approach.

In block detection stage, there are mainly three types of errors. The first type of errors was due to extra line breaks in the text, which mistakenly breaks the targeted instance into multiple lines. The second type of errors was because of extra spaces in the Chinese text (note space in the Chinese text space is different from that in the English text), e.g. "上海市零陵路" is mistakenly written as "上海 市零陵路".

In text annotation stage, errors can be summarized into two categories. The first type of errors was due to the errors at the block detection step. The second type of errors was due to errors of detection of instances' end position.

6. Conclusion

In this paper, we have investigated the problem of semantic annotation using horizontal and vertical context. We propose a two-stage approach on the basis of machine learning methods. The proposed approach has been applied to annotate company annual reports. Experimental results show that our approach can significantly outperform the baseline methods as well as the sequence-base methods.

References

[1] R. Benjamins and J. Contreras. Six Challenges for the Semantic Web. Intelligent Software Components. Intelligent software for the networked economy (isoco). April, 2002

[2] N. Kushmerick, D.S. Weld, and R.B. Doorenbos. Wrapper Induction for Information Extraction. In Proc. of IJCAI. Nagoya, Japan. 1997:729-737

[3] F. Ciravegna. (LP)2, an Adaptive Algorithm for Information Extraction from Web-related Texts. In Proc. of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, Seattle, USA. August 2001

[4] S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM—Semi-automatic Creation of Metadata, In Proc. of EKAW 2002, Siguenza, Spain, 2002: 358-372

[5] J. Tang, J. Li, H. Lu, B. Liang, and K. Wang. iASA: Learning to Annotate the Semantic Web. Journal on Data Semantic. 2005, Vol(4): 110-145

[6] C. Cortes and V. Vapnik. Support-Vector Networks. Machine Learning, Vol(20), pp273-297. 1995

[7] L. Reeve. Integrating Hidden Markov Models into Semantic Web Annotation Platforms. Technique Report. 2004