Unsupervised Information Extraction by Text Segmentation

Eli Cortez, Altigran Soares da Silva (Orientador)

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brazil

eccv@dcc.ufam.edu.br alti@icomp.ufam.edu.br

Abstract. *In this work we propose, implement and evaluate a new unsupervised approach for the problem of Information Extraction by Text Segmentation (IETS). Our approach relies on information available on pre-existing data to learn how to associate segments in the input with attributes of a given domain relying on a very effective set of content-based features. The effectiveness of these content-based features is also exploited to directly learn from test data structure-based features, with no previous human-driven training, a feature unique to our approach. Based on our approach, we have produced a number of results to address the IETS problem. We have performed different experiments that indicate that our approach yields high quality results when compared to state-of-the-art approaches and that it is able to properly support IETS methods in a number of real applications.*

Resumo. *Neste trabalho, propômos, implementamos e avaliamos uma nova abordagem não-supervisionada para o problema de Extração de Informação por Segmentação de Texto (EIST). Nossa abordagem baseia-se em informações disponíveis em dados pré-existentes para aprender a associar segmentos de texto com atributos de um determinado domínio utilizando um conjunto muito eficaz de características baseadas em conteúdo. A eficácia das características baseadas em conteúdo também é explorada para aprender diretamente dos textos de entrada características baseadas em estrutura, sem nenhuma intervenção humana, uma característica única da nossa abordagem. Com base em nossa abordagem, produzimos inúmeros resultados para lidar com problema de EIST. Nós realizamos diferentes experimentos que indicam que a nossa abordagem produz resultados de alta qualidade em relação ao estado-da-arte e que é capaz de amparar adequadamente métodos de EIST em uma série de aplicações reais.*

1. Introduction

Over the last years, there has been a steady increase in the number and types of sources of textual information in the World-Wide Web. Examples of such sources are e-shops, digital libraries, social networks, etc. In most cases, these sources are freely accessible, cover a variety of topics and subjects, provide information in distinct formats and styles, and do not impose any rigid publication format. In addition, they are constantly kept up-to-date by users and organizations. Through the eyes of data management scientists, these sources constitute large repositories of valuable data on a variety of domains. Depending on the type of each source, one can find in them data referring to personal information, products, publications, companies, cities, weather, etc., from which it is possible to perform tasks, such as to infer relationships, to learn user preferences and to detect trends, etc.

Nevertheless, the abundance and popularity of these online sources of relevant data have attracted a number of research efforts to address problems related to them, such as crawling,

extracting, querying and mining. In particular, the extraction problem, commonly know as *Information Extraction (IE)* in the literature [Sarawagi 2008], refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from noisy unstructured sources. This problem is the main subject of this work.

The IE problem encompasses distinct sub-problems, including the Information Extraction by Text Segmentation (IETS) problem. IETS is the problem of segmenting unstructured textual inputs to extract implicit data values contained in them. To better illustrate this problem, consider Figure 1. Figure 1(a) depicts a unstructured record (postal address). This record contains relevant information such as: person name, street name, house number, zip code, etc., and does not contain any explicit delimiter between the values composing it. Figure 1(b) shows an expected output for this example, where each segment receives a label indicating that the text segment contains a value of an attribute.

Figure 1. Unstructured textual record (a) and an expected output (b).

Considering the practical and theoretical importance of the IETS problem [Borkar et al. 2001, Sarawagi 2008, Zhao et al. 2008], we proposed, implemented and evaluated an unsupervised approach to address it. Our approach relies on pre-existing data in the form of a *knowledge base*, to provide features for a learning process.

Based on a knowledge base, our approach learns how to associate segments in the input string with attributes of a given domain relying on a very effective set of content-based features, which characterize the domain of the attributes (e.g., typical values, terms composing them, their format, etc.). The effectiveness of these content-based features is also exploited to directly learn from test data structure-based features (e.g., the positioning and sequencing of attribute values, etc.), which characterize the structure of the records within the source text., with no previous human-driven training, a characteristic that is unique to our approach.

Based on our approach, we have produced a number of results to address the IETS problem in a unsupervised fashion. Particularly, we have developed, implemented and evaluated distinct state-of-the-art IETS methods which were published at top tier venues, such as: SIGMOD Conference, VLDB Conference, WWW Conference and JASIST.

- For the case where the input unstructured records are explicitly delimited, we propose a method called *ONDUX* [Cortez et al. 2010b, Cortez et al. 2010a, Porto et al. 2011][1].
- For dealing with textual inputs that do not contain any explicit structural information available, we have developed a method called *JUDIE* [Cortez et al. 2011a].
- We have also developed a method, called *iForm*, that relies on our extraction approach to perform the task of Web form filling [Toda et al. 2009, Toda et al. 2010].

2. Related Work

In the literature, different approaches have been proposed to address the Information Extraction by Text Segmentation problem. The first proposed approaches [Borkar et al. 2001, Sarawagi 2008] were all based on supervised techniques that heavily rely on manually created training instances. Recent approaches presented in the literature propose the use of pre-existing data for easing the training process [Mansuri and Sarawagi 2006, Zhao et al. 2008]. These approaches take advantage of the existence of large amounts of structured datasets that can be used with little or no user effort. The main differences between our proposed approach and the previous approaches are: (1) the way content-based features are learned; (2) how structure-based features are automatically induced and (3) the way our approach automatically combines

[1]Tool awarded as the Best Tool of the Symposium

these features in a fully automatic fashion. A detailed discussion about related work can be found in [Cortez 2012].

3. Exploiting Pre-Existing Datasets to Support IETS

In this section we present our proposed unsupervised approach for the IETS problem. More details are presented in [Cortez 2012]. Consider a set of data-rich input text snippets from which we need to extract data contained in them. We assume that all snippets in this set belong to the same application domain (e.g., product descriptions, bibliographic citations, postal addresses, etc). We also assume the existence of a dataset on the same domain as the input set, which we call *Knowledge Base*. A Knowledge Base is a set of pairs $K = \{\langle A_1, O_1\rangle, \ldots, \langle A_n, O_n\rangle\}$ in which each A_i is a distinct attribute and O_i is a set of strings $\{o_{i,1}, \ldots, o_{i,n_i}\}$ called *occurrences*. Intuitively, O_i is a set of strings representing plausible or typical values for an attribute A_i. In Figure 2 we illustrate a very simple example of a knowledge base.

$$
\begin{aligned}
K =& \{\langle Neigh., O_{Neigh.}\rangle, \langle Street, O_{Street}\rangle, \langle Bathrooms, O_{Bathrooms}\rangle\} \\
O_{Neighborhood} =& \{\text{“}\textbf{Regent Square}\text{”},\text{“}\textbf{Milenight Park}\text{”}\} \\
O_{Street} =& \{\text{“}\textbf{Regent St.}\text{”},\text{“}\textbf{Morewood Ave.}\text{”},\text{“}\textbf{Square Ave. Park}\text{”}\} \\
O_{Bathrooms} =& \{\text{“}\textbf{Two Bathrooms}\text{”},\text{“}\textbf{5 Bathrooms}\text{”}\}
\end{aligned}
$$

Figure 2. A simple example of a Knowledge Base.

Our proposed approach to tackle the information extraction by text segmentation problem, relies on the following steps, which are illustrated in Figure 3: (1) learn content-based features (g^k) from a knowledge base, (2) use the learned content-based features in an initial extraction process, (3) explore the outcome of the initial extraction process to automatically induce structure-based features (f^k) and (4) combine content-based features with structure-based features to achieve a final extraction result. Thus, our proposed approach relies on the hypothesis that the usage of knowledge bases allow for the unsupervised learning of both content-based and structure-based features.

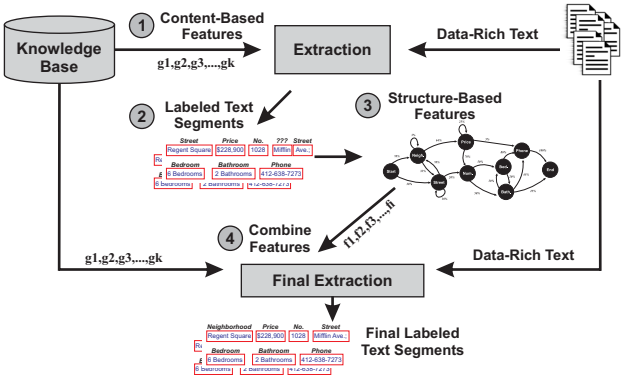

Figure 3. Overview of our proposed approach.

Consider a candidate unstructured record $R = s_1, \ldots, s_r$, where each text segment $s_i (1 \leq i \leq r)$ is a candidate value. Also, consider an attribute A and let ℓ_A be a label used for this attribute. Then, for any text segment s_i, we can compute the value of the content-based feature $g^k(s_i, A, R)$ and the structure-based feature $f^k(s_i, A, R)$. Function g^k returns a real number that measures how well a hypothetical value formed by tokens in the text segment s_i follows some property of the values in the domain of A. Function f^k returns a real number that measures the likelihood of a segment labeled ℓ_A to occur in the same place as s_i in R.

Different content-based features can be learned from the knowledge encoded in the knowledge bases, which are exploited by our approach. These features are: *attribute vocabulary*, *attribute value range* and *attribute value format*. A very important point regarding these features is the fact that they can be computed from previously available knowledge bases and, thus, they are independent of the target text corpus, that is, these features are *input-independent*.

$$AF(s,A) = \frac{\sum_{t \in T(A) \cap T(s)} \textit{fitness}(t,A)}{|T(s)|} \text{ where, } fitness(t,A) = \frac{f(t,A)}{N(t)} \times \frac{f(t,A)}{f_{max}(A)} \quad (1)$$

The attribute vocabulary feature, implemented in our *AF function* (Eq. 1), exploits the common terms often shared by values of textual attributes. The attribute value range feature, implemented in our *NM function* (Eq. 2), specifically deals with numeric attributes using the average and the standard deviation of the values of numeric attributes available on the knowledge base. Finally, the attribute value format feature, implemented in our *format function* (Eq. 3), exploits the writing styles often used to represent values of different attributes in the knowledge base (e.g., url, date, telephone). We assume that these features exploit different properties of the attribute domain, thus, we can say they are independent, what allows us to combine them by means of the Bayesian disjunctive operator *or*, also known as *Noisy-OR-Gate*.

$$NM(s,A) = e^{-\frac{v_s - \mu_A}{2\sigma_A^2}} \quad (2) \qquad format(s,A) = \frac{\sum_{\langle n_x, n_y \rangle \in path(s)} w(n_x, n_y)}{|path(s)|} \quad (3)$$

Experiments we have performed on a number of different scenarios and with datasets in distinct domains indicated that our approach is able to achieve good extraction quality relying only on content-based features. However, there are cases in which we can further exploit these features to automatically induce structure-based features and improve the quality of the extraction results. For computing such structure-based features, it is common to use a graph model that represents the likelihood of attribute transitions within the input text (or any other input text from the same source). We use a probabilistic HMM-like graph model that we call PSM (Positioning and Sequencing Model). With the structure-based features in hand, we can use them to improve the initial extraction that resorted only on content-based features. Now, taking into consideration content-based and structure-based feature, function $\ell(s,R,A)$ (Eq. 4) is computed, for each candidate segment s of all candidate records R in the input text, for all attributes A of the same data type (i.e., text or numeric). Thus, s is labeled with a label representing the attribute that yielded the highest score according to ℓ.

$$\begin{aligned} \ell(s,R,A) = & 1 - ((1 - g^1(s,A)) \times \ldots \times (1 - g^n(s,A)) \times \\ & (1 - f^1(s,A,R)) \times \ldots \times (1 - f^m(s,A,R))) \end{aligned} \quad (4)$$

4. Proposed State-of-the-Art Methods based on our Unsupervised Approach

Based on our approach, we have developed, implemented and evaluated distinct state-of-the-art IETS methods. In order to evaluate the performance of these methods we have performed different experiments mostly using publicly available datasets[2]. The quality of the extraction tasks was measured using precision, recall and F-Measure. In here, we present only a summary of each method and its main results. The description of each method and its detailed experimental evaluation, including the details of baselines can be found in [Cortez 2012].

For the case where the input unstructured records are explicitly delimited in the input text, we propose a method called *ONDUX* [Cortez et al. 2010b, Cortez et al. 2010a, Porto et al. 2011]. *ONDUX* (On Demand Unsupervised Information Extraction) is an unsupervised probabilistic method for IETS. Like other unsupervised IETS methods, *ONDUX* relies on information available on pre-existing data, but, unlike previously proposed methods, it also

[2] http://www.isi.edu/integration/RISE/

relies on a very effective set of content-based features to bootstrap the learning of structure-based features. More specifically, structure-based features are exploited to disambiguate the extraction of certain attributes through a reinforcement step. The novel reinforcement step relies on sequencing and positioning of attribute values directly learned *on-demand* from test data. This assigns to *ONDUX* a high degree of flexibility and considerably improves its effectiveness, as demonstrated by the experimental evaluation with textual sources from different domains. In our experimental evaluation, *ONDUX* achieved the best results in all datasets in terms of attribute-level extraction, providing, on average, **0.92** of F-Measure while the U-CRF [Zhao et al. 2008] baseline achieved 0.78 of F-Measure. Real applications use *ONDUX* to perform IE tasks, one example is Ciência Brasil[3] [Laender et al. 2011], a social network for scientists. More details of *ONDUX* are available in [Cortez 2012], page 37.

We have also developed a method called *JUDIE* [Cortez et al. 2011a], for dealing with textual inputs that do not contain any explicit structural information available. *JUDIE* (Joint Unsupervised Structure Discovery and Information Extraction) is a method for automatically extracting semi-structured data records in the form of continuous text (e.g., bibliographic citations, postal addresses, etc.) and having no explicit delimiters between them. *JUDIE* is capable of detecting the structure of each individual record being extracted without any user assistance. This is accomplished by a novel Structure Discovery algorithm that, given a sequence of labels representing attributes assigned to potential values, groups these labels into individual records by looking for frequent patterns of label repetitions in the given sequence. In comparison with other methods, including *ONDUX*, *JUDIE* faces a task considerably harder, that is, extracting information while simultaneously uncovering the underlying structure of the implicit records containing it. In our experimental evaluation, *JUDIE* achieved better results in all datasets in terms of attribute-level extraction, providing, on average, **0.89** of F-Measure while the U-CRF baseline achieved 0.73 of F-Measure. More details are available in [Cortez 2012], page 63.

We also rely on our proposed approach to develop a method, called *iForm*, for dealing with the task of Web form filling [Toda et al. 2009, Toda et al. 2010]. In this case, the aim is at extracting segments from a data-rich text given as input and associating these segments with fields from a target Web form. The extraction process relies on content-based features learned from data that was previously submitted to the Web form. In order to evaluate *iForm* we have used real datasets. Some of these datasets were taken from the auction web site *TodaOferta*[4]. On average, *iForm* was able to achieve more than **0.71** of F-Measure considering different multi-typed web forms. In comparison with the baseline, iCRF, *iForm* achieved better form filling quality. These experiments corroborate our claims that the usage our approach in *iForm* is feasible and effective, and that it works well even when only a few previous submissions to the input interface are available. More details are available in [Cortez 2012], page 91.

5. Conclusions and Future Work

In this paper we summarized some of the research contributions achieved in our doctorate work [Cortez 2012]. Specifically, we have proposed, implemented and evaluated an unsupervised approach for the problem of Information Extraction by Text Segmentation (IETS).

We have studied different aspects regarding our approach and compared it with state-of-the-art IE methods. Results indicate that our approach performs quite well when compared with such methods, even without any user intervention. Based on our approach, we have produced a number of results to address the IETS problem in a unsupervised fashion. Our state-of-the-art

[3] http://www.pbct.inweb.org.br/pbct/
[4] http://todaoferta.uol.com.br/

IETS methods were published in top tier venues. The papers produced during this PhD were published in 4 of the top 6 venues of the Databases & Information Systems area according to the rank provided by Google Scholar[5]. Also, according to the Qualis (CAPES), we have published 6 papers in venues A1, 1 paper in venues A2 and 3 papers in venues B3. The results achieved during this PhD were present in tutorials of national and international venues. Finally, we notice that our techniques are in use in a startup company called Neemu Technologies[6], which currently holds the largest market share in search systems for e-commerce in Brazil.

Due to lack of space, we were not able to include here further results such as a strategy for automatically obtaining knowledge bases from the Wikipedia [Serra et al. 2011] and a method for IETS in bibliographic references [Cortez et al. 2009], as well as other web data management results [Cortez et al. 2011b], [Evangelista et al. 2010], [Evangelista et al. 2009][7].

References

Borkar, Deshmukh, and Sarawagi (2001). Automatic Segmentation of Text into Structured Records. In *Int. Conf. on Manag. of Data (SIGMOD)*, pages 175–186.

Cortez (2012). *Unsupervised Approach for Information Extraction by Text Segmentation*. Phd thesis, Universidade Federal do Amazonas.

Cortez, et al. (2010a). ONDUX: On-Demand Unsupervised Learning for Information Extraction. In *Int. Conf. on Manag. of Data (SIGMOD)*, pages 807–818.

Cortez, et al. (2011a). Joint unsupervised structure discovery and information extraction. In *Int. Conf. on Manag. of Data (SIGMOD)*, pages 541–552.

Cortez, et al. (2009). A flexible approach for extracting metadata from bibliographic citations. *J. American Soc. for Inf. Science and Tech. (JASIST)*, 60(6):1144–1158.

Cortez, et al. (2010b). Unsupervised strategies for information extraction by text segmentation. In *SIGMOD PhD Workshop on Innov. Database Res.*, pages 49–54.

Cortez, et al. (2011b). Lightweight methods for large-scale product categorization. *J. American Soc. for Inf. Science and Tech. (JASIST)*, 62(9):1839–1848.

Evangelista, et al. (2010). Adaptive and flexible blocking for record linkage tasks. *J. Inf. and Data Manag. (JIDM)*, 1(2):167.

Evangelista, et al. (2009). Blocagem adaptativa e flexível para o pareamento aproximado de registros. In *Simp. Bras. de Banco de Dados (SBBD)*, pages 61–75.

Laender, et al. (2011). Building a research social network from an individual perspective. In *Joint Conf. on Dig. Libraries (JCDL)*, pages 427–428.

Mansuri and Sarawagi (2006). Integrating Unstructured Data into Relational Databases. In *Int. Conf. on Data Engineering (ICDE)*, pages 29–41.

Porto, et al. (2011). Unsupervised information extraction with the ondux tool. In *Simp. Bras. de Banco de Dados (SBBD)*.

Sarawagi (2008). Information extraction. *Found. Trends in Databases*, 1(3):261–377.

Serra, et al. (2011). On using wikipedia to build knowledge bases for information extraction by text segmentation. *J. Inf. and Data Manag. (JIDM)*, 2(3):259.

Toda, et al. (2010). A probabilistic approach for automatically filling form-based web interfaces. *Proceedings of the VLDB Endowment*, 4(3):151–160.

Toda, et al. (2009). Automatically filling form-based web interfaces with free text inputs. In *Int. World Wide Web Conf. (WWW)*, pages 1163–1164.

Zhao, et al. (2008). Exploiting structured reference data for unsupervised text segmentation with conditional random fields. In *SIAM Int. Conf. on Data Min.*, pages 420–431.

[5] http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_databasesinformationsystems

[6] http://neemu.com/site/

[7] Paper awarded as the Best Paper of the Symposium