Using Conditional Random Fields to Exploit Token Structure and Labels for Accurate Semantic Annotation

Aman Goel *, **Craig A. Knoblock** and **Kristina Lerman**
Information Sciences Institute and Computer Science Department
University of Southern California
4676, Admiralty Way
Marina del Rey, CA 90292
*amangoel@isi.edu

Abstract

Automatic semantic annotation of structured data enables unsupervised integration of data from heterogeneous sources but is difficult to perform accurately due to the presence of many numeric fields and proper-noun fields that do not allow reference-based approaches and the absence of natural language text that prevents the use of language-based approaches. In addition, several of these semantic types have multiple heterogeneous representations, while sharing syntactic structure with other types. In this work, we propose a new approach to use conditional random fields (CRFs) to perform semantic annotation of structured data that takes advantage of the structure and labels of the tokens for higher accuracy of field labeling, while still allowing the use of exact inference techniques. We compare our approach with a linear-CRF based model that only labels fields and also with a regular-expression based approach.

Introduction

Semantic annotation is the problem of assigning user-defined semantic labels to fields and tokens in structured data. It allows automatic joining of different sources on common attributes by identifying and comparing their values even when they are in different syntactic formats. Automatic semantic annotation is difficult due to heterogeneity in the formats of semantic types, as well as similarity between different semantic types. For example, *Temperature* can be written as $56°F$, $56\ F$, $56°$, or 56, whereas *Humidity* and *Chance of Precipitation* look very similar (e.g., 40%).

Conventional language-based methods cannot be applied to this problem due to the lack of well formed sentences and the presence of many numeric and string literal values prevents the use of a reference-set. The primary evidence of a semantic type is its token-level syntactic structure and the identity of the neighboring fields. Our contribution in this paper is an approach to perform semantic annotation of structured data using CRFs (Lafferty, McCallum, and Pereira 2001) that take advantage of the label interdependencies between a field label and its token's labels, the label interdependencies between consecutive tokens, and the dependence of these labels on the features of the tokens to achieve higher field labeling accuracy, while still allowing us to use exact inference techniques.

Table 1: A sample data tuple, its fields, tokens and semantic labels.

Fields	Field labels	Tokens	Token labels
90292	Zip	90292	ZipValue
$76°F$	TempF	76	TempFValue
		°	DegreeSymbol
		F	TempFUnit
50%	Humidity	50	HumidityValue
		$\%$	PercentSymbol
$5mph$	WindSpeed	5	WindSpeedValue
		mph	WindSpeedUnit

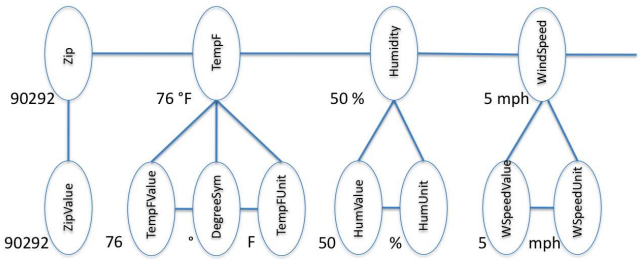

Figure 1: CRF graph generated for data tuple shown in Table 1.

Semantic annotation using a CRF-based model

We automatically construct the CRF graphs (e.g., Figure 1) from tuples (e.g., Table 1) using our algorithm, which is as follows: Construct a node for each field and connect them in a chain. Split each field into tokens and add nodes corresponding to them as children to the field node. Connect neighboring token nodes. Our parser tokenizes the fields into continuous strings of alphabets, pure numbers and single symbol characters as shown in the example tuple.

*Supported by IARPA via AFRL contract number FA8650-10-C-7058. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

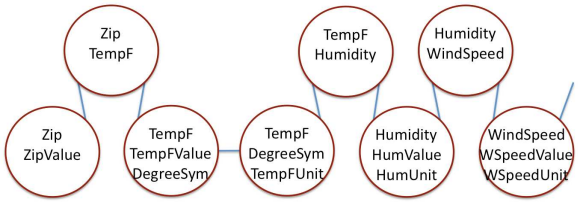

Figure 2: Junction tree constructed from CRF graph in Figure 1.

We then attach some of the following syntactic features to each token: capitalization, length of token, starting character or digit, number of digits after decimal, digit at units place, sign of the number, and identity of the token itself.

We use seven types of feature functions to capture the relationships between the following: (1) the label and feature on a token node, (2) the label of a token node and feature of its previous token, (3) the label of a token node and feature of its next token, (4) the label of a field node and presence of a feature on any of its tokens, (5) the labels of two adjacent token nodes and their features, (6) the labels of a field and its token node and features of the token, and (7) the labels of two adjacent field nodes and presence of a feature each on any of their respective tokens. We use binary feature functions, that return either zero or one. For example, a feature function of the third type defined above can take as inputs the label of a token node and the feature list of next token and return the value one only if the first node is labeled *HouseNumber* and the next token has the feature *Capitalized*.

Training of a model and labeling of a new tuple, both involve performing inference on the CRFs. Since our CRFs have loops, we use the junction tree algorithm (Lauritzen and Spiegelhalter 1988) to convert our CRF graphs into acyclic junction trees (JT) (Figure 2). Each node in the JT represents a clique in the CRF graph. Figure 2 shows the JT for the CRF graph shown in Figure 1. It is the property of our CRF graph structure that it leads to a linear chain JT, which makes using belief propagation (BP) (Pearl 1988) on it very easy. Since BP calculates the beliefs for all possible label assignments and one node of JT represents a clique in the CRF, the maximum number of label assignments to a JT node is exponential in the size of the largest clique, which is three in our CRFs. This keeps the complexity low and avoids the need for approximate inference methods.

Experiments

We collected data by extracting 15 tuples each from four websites in three domains. In each domain, we ran four experiments, each time training on data from three websites and testing on the fourth website. The details of the domains and the results are reported in Table 2. Field labeling accuracy is more than 80% for 10 out of the 12 websites and token labeling accuracy is around 80% for all the three domains. The average accuracy across all domains is 88%. We compared our performance with a regular-expression-based approach and a linear-CRF-based approach, where the fields are not split into tokens and found their accuracy to be 75% and 48%, respectively.

Table 2: Performance of our approach on 12 websites.

Domain (field types, token types)	**Website URL**	**Field labeling accuracy**	**Token labeling accuracy**
Weather forecast (15, 36)	wunderground.com	0.89	0.92
	weather.unisys.com	0.43	0.75
	weather.com	0.70	0.79
	noaa.gov	1.00	0.86
	average	**0.75**	**0.83**
Flight status (8, 17)	flytecomm.com	0.89	0.82
	flightview.com	0.96	0.97
	delta.com	0.81	0.78
	continental.com	0.96	0.55
	average	**0.90**	**0.79**
Geocoding (5,12)	geocoder.us	1.00	0.85
	geocoder.ca	1.00	0.82
	geonames.com	0.98	0.68
	worldkit.com	1.00	0.89
	average	**0.99**	**0.81**

Related Work

Zhu et al. (2005) used two-dimensional CRFs and (Tang et al. 2006) used tree-structured CRFs to exploit spatial relationships between elements on webpages and documents, respectively. Our hierarchical-CRF exploits the semantic relationship between field and token labels in data tuples.

Schema matching techniques (Doan, Domingos, and Levy 2000) can be used to perform semantic annotation by column-wise matching between labeled and unlabeled relational tables. These techniques assume that there are a large number of rows in each table. Although we train our model on multiple tuples, we only label one tuple at a time. This allows us to label variable length tuples.

Conclusion

In this paper, we presented a CRF-based approach to exploit the token-level structure of the fields to perform accurate semantic annotation of structured data. We showed that this approach achieves about 13% higher accuracy than a linear-CRF model that assigns labels to fields only.

References

Doan, A.; Domingos, P.; and Levy, A. Y. 2000. Learning source descriptions for data integration. *WebDB* 81-86.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the ICML*, 282–289.

Lauritzen, S. L., and Spiegelhalter, D. J. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society* 50(2).

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 2nd edition.

Tang, J.; Hong, M.; Li, J.; and Liang, B. 2006. Tree-structured conditional random fields for semantic annotation. In *ISWC*.

Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions*.

Zhu, J.; Nie, Z.; Wen, J.; Zhang, B.; and Ma, W. 2005. 2d conditional random fields for web information extraction. In 22^{nd} *ICML*.