Pattern Learning for Relation Extraction with a Hierarchical Topic Model

Enrique Alfonseca Katja Filippova Jean-Yves Delort
Google Research
Brandschenkestrasse 110
8002 Zurich, Switzerland
{ealfonseca,katjaf,jydelort}@google.com

Guillermo Garrido[*]
NLP & IR Group, UNED
Juan del Rosal, 16.
28040 Madrid, Spain
ggarrido@lsi.uned.es

Abstract

We describe the use of a hierarchical topic model for automatically identifying syntactic and lexical patterns that explicitly state ontological relations. We leverage distant supervision using relations from the knowledge base FreeBase, but do not require any manual heuristic nor manual seed list selections. Results show that the learned patterns can be used to extract new relations with good precision.

1 Introduction

The detection of relations between entities for the automatic population of knowledge bases is very useful for solving tasks such as Entity Disambiguation, Information Retrieval and Question Answering. The availability of high-coverage, general-purpose knowledge bases enable the automatic identification and disambiguation of entities in text and its applications (Bunescu and Pasca, 2006; Cucerzan, 2007; McNamee and Dang, 2009; Kwok et al., 2001; Pasca et al., 2006; Weld et al., 2008; Pereira et al., 2009; Kasneci et al., 2009).

Most early works in this area were designed for supervised Information Extraction competitions such as MUC (Sundheim and Chinchor, 1993) and ACE (ACE, 2004; Doddington et al., 2004; Li et al., 2011), which rely on the availability of annotated data. Open Information Extraction (Sekine, 2006; Banko et al., 2007; Bollegala et al., 2010) started as an effort to approach relation extraction in a completely unsupervised way, by learning regularities and patterns from the web. Two example systems implementing this paradigm are TEXTRUNNER (Yates et al., 2007) and REVERB (Fader et al., 2011). These systems do not need any manual data or rules, but the relational facts they extract are not immediately disambiguated to entities and relations from a knowledge base.

A different family of unsupervised methods for relation extraction is *unsupervised semantic parsing*, which aims at clustering entity mentions and relation surface forms, thus generating a semantic representation of the texts on which inference may be used. Some techniques that have been used are Markov Random Fields (Poon and Domingos, 2009) and Bayesian generative models (Titov and Klementiev, 2011). These are quite powerful approaches but have very high computational requirements (cf. (Yao et al., 2011)).

A good trade-off between fully supervised and fully unsupervised approaches is *distant supervision*, a semi-supervised procedure consisting of finding sentences that contain two entities whose relation we know, and using those sentences as training examples for a supervised classifier (Hoffmann et al., 2010; Wu and Weld, 2010; Hoffmann et al., 2011; Wang et al., 2011; Yao et al., 2011). A usual problem is that two related entities may co-occur in one sentence for many unrelated reasons. For example, *Barack Obama* is the president of the *United States*, but not every sentence including the two entities supports and states this relation. Much of the previous work uses heuristics, e.g. extracting sentences only from encyclopedic entries (Mintz et al.,

[*] Work done during an internship at Google Zurich.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 54–59,
Jeju, Republic of Korea, 8-14 July 2012.

2009; Hoffmann et al., 2011; Wang et al., 2011), or syntactic restrictions on the sentences and the entity mentions (Wu and Weld, 2010). These are usually defined manually and may need to be adapted to different languages and domains. Manually selected seeds can also be used (Ravichandran and Hovy, 2002; Kozareva and Hovy, 2010).

The main contribution of this work is presenting a variant of *distance supervision* for relation extraction where we do not use heuristics in the selection of the training data. Instead, we use topic models to discriminate between the patterns that are expressing the relation and those that are ambiguous and can be applied across relations. In this way, high-precision extraction patterns can be learned without the need of any manual intervention.

2 Unsupervised relational pattern learning

Similar to other distant supervision methods, our approach takes as input an existing knowledge base containing entities and relations, and a textual corpus. In this work it is not necessary for the corpus to be related to the knowledge base. In what follows we assume that all the relations studied are binary and hold between exactly two entities in the knowledge base. We also assume a dependency parser is available, and that the entities have been automatically disambiguated using the knowledge base as sense inventory.

One of the most important problems to solve in distant supervision approaches is to be able to distinguish which of the textual examples that include two related entities, e_i and e_j, are supporting the relation. This section describes a fully unsupervised solution to this problem, computing the probability that a pattern supports a given relation, which will allow us to determine the most likely relation expressed in any sentence. Specifically, if a sentence contains two entities, e_i and e_j, connected through a pattern w, our model computes the probability that the pattern is expressing any relation –$P(r|w)$– for any relation r defined in the knowledge base. Note that we refer to patterns with the symbol w, as they are the words in our topic models.

Preprocessing As a first step, the textual corpus is processed and the data is transformed in the following way: (a) the input corpus is parsed and entities are disambiguated; (b) for each relation r in the knowledge base, a new (initially empty) document collection C_r is created; (c) for each entity pair (e_i, e_j) which are related in the knowledge base, a new (initially empty) document D_{ij} is created; (d) for each sentence in the input corpus containing one mention of e_i and one mention of e_j, a new term is added to D_{ij} consisting of the context in which the two entities were seen in the document. This context may be a complex structure, such as the dependency path joining the two entities, but it is considered for our purposes as a single term; (e) for each relation r relating e_i with e_j, document D_{ij} is added to collection C_r. Note that if the two entities are related in different ways at the same time, an identical copy of the document D_{ij} will be added to the collection for all those relations.

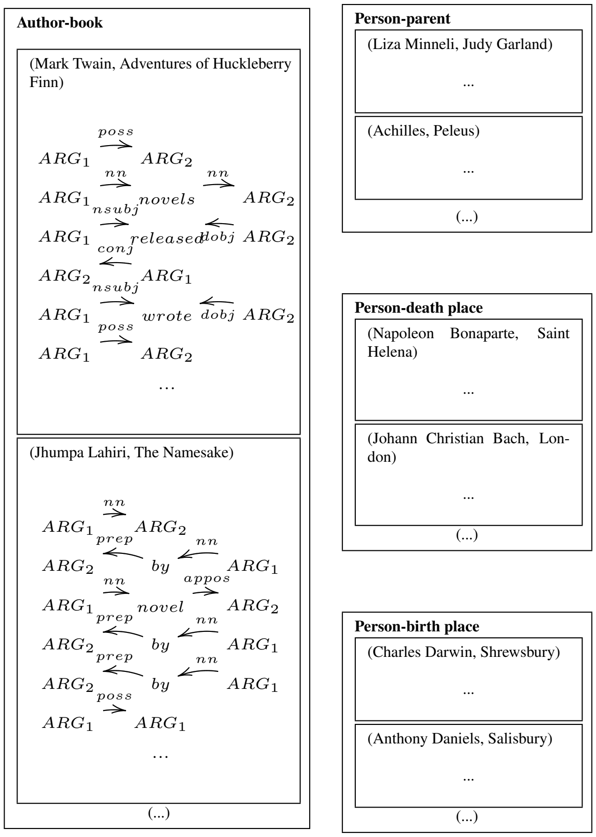

Author-book
(Mark Twain, Adventures of Huckleberry Finn)
$ARG_1 \xrightarrow{poss} ARG_2$
$ARG_1 \xrightarrow{nn} novels \xrightarrow{nn} ARG_2$
$ARG_1 \xrightarrow{nsubj} released \xleftarrow{dobj} ARG_2$
$ARG_2 \xleftarrow{conj} ARG_1$
$ARG_1 \xrightarrow{nsubj} wrote \xleftarrow{dobj} ARG_2$
$ARG_1 \xrightarrow{poss} ARG_2$
...
(Jhumpa Lahiri, The Namesake)
$ARG_1 \xrightarrow{nn} ARG_2$
$ARG_2 \xleftarrow{prep} by \xleftarrow{nn} ARG_1$
$ARG_1 \xrightarrow{nn} novel \xrightarrow{appos} ARG_2$
$ARG_2 \xleftarrow{prep} by \xleftarrow{nn} ARG_1$
$ARG_2 \xleftarrow{prep} by \xleftarrow{nn} ARG_1$
$ARG_1 \xrightarrow{poss} ARG_1$
...
(...)
Person-parent
(Liza Minneli, Judy Garland)
...
(Achilles, Peleus)
...
(...)
Person-death place
(Napoleon Bonaparte, Saint Helena)
...
(Johann Christian Bach, London)
...
(...)
Person-birth place
(Charles Darwin, Shrewsbury)
...
(Anthony Daniels, Salisbury)
...
(...)

Figure 1: Example of a generated set of document collections from a news corpus for relation extraction. Larger boxes are document collections (relations), and inner boxes are documents (entity pairs). Document contain dependency patterns, which are *words* in the topic model.

Figure 1 shows a set of document collections gen-

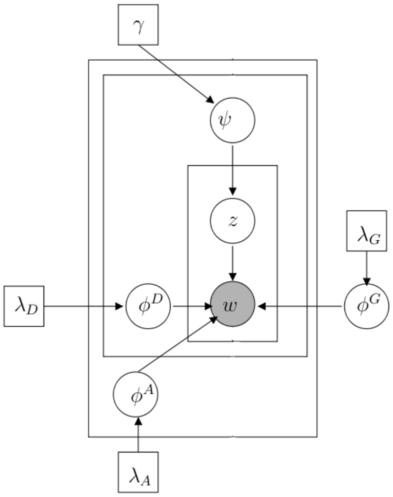


Figure 2: Plate diagram of the generative model used.

erated for three relations using this procedure. Each relation r has associated a different document collection, which contains one document associated to each entity pair from the knowledge base which is in relation r. The *words* in each document can be, for example, all the dependency paths that have been observed in the input textual corpus between the two related entities. Each document will contain some very generic paths (e.g. the two entities consecutive in the text) and some more specific paths.

Generative model Once these collections are built, we use the generative model from Figure 2 to learn the probability that a dependency path is conveying some relation between the entities it connects. This model is very similar to the one used by Haghighi and Vanderwende (2009) in the context of text summarization. w (the observed variable) represents a pattern between two entities. The topic model ϕ^G captures general patterns that appear for all relations. ϕ^D captures patterns that are specific about a certain entity pair, but which are not generalizable across all pairs with the same relation. Finally ϕ^A contains the patterns that are observed across most pairs related with the same relation. ϕ^A is the topic model of interest for us.

We use Gibbs sampling to estimate the different models from the source data. The topic assignments (for each pattern) that are the output of this process are used to estimate $P(r|w)$: when we observe pattern w, the probability that it conveys relation r.

3 Experiments and results

Settings We use Freebase as our knowledge base. It can be freely downloaded[1]. text corpus used contains 33 million English news articles that we downloaded between January 2004 and December 2011. A random sample of 3M of them is used for building the document collections on which to train the topic models, and the remaining 30M is used for testing. The corpus is preprocessed by identifying Freebase entity mentions, using an approach similar to (Milne and Witten, 2008), and parsing it with an inductive dependency parser (Nivre, 2006).

From the three million training documents, a set of document collections (one per relation) has been generated, by considering the sentences that contain two entities which are related in FreeBase through any binary relation and restricting to high-frequency 200 relations. Two ways of extracting patterns have been used: (a) **Syntactic**, taking the dependency path between the two entities, and (b) **Intertext**, taking the text between the two. In both cases, a topic model has been trained to learn the probability of a relation given a pattern w: $p(r|w)$. For λ we use symmetric Dirichlet priors $\lambda_G = 0.1$ and $\lambda_D = \lambda_{\mathcal{A}} = 0.001$, following the intuition that for the background the probability mass across patterns should be more evenly distributed. γ is set as (15, 15, 1), indicating in the prior that we expect more patterns to belong to the background and entity-pair-specific distributions due to the very noisy nature of the input data. These values have not been tuned.

As a baseline, using the same training corpus, we have calculated $p(r|w)$ using the maximum likelihood estimate: the number of times that a pattern w has been seen connecting two entities for which r holds divided by the total frequency of the pattern.

Extractions evaluation The patterns have been applied to the 30 million documents left for testing. For each pair of entities disambiguated as FreeBase entities, if they are connected through a known pattern, they are assigned $\arg\max_r p(r|w)$. We have randomly sampled 4,000 such extractions and sent them to raters. An extraction is to be judged correct if both it is correct in real life and the sentence from which it was extracted really supports it. We

[1] `http://wiki.freebase.com/wiki/Data_dumps`

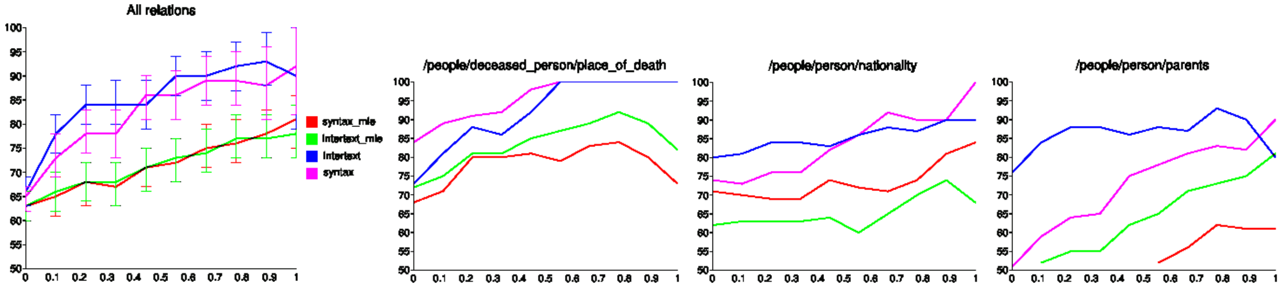


Figure 3: Evaluation of the extractions. X-axis has the threshold for $p(r|w)$, and Y-axis has the precision of the extractions as a percentage.

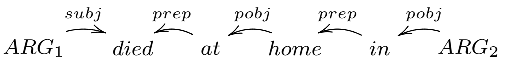
have collected three ratings per example and taken the majority decision. There was disagreement for 9.4% of the items on whether the sentence supports the relation, and for 20% of the items on whether the relation holds in the real world.

The results for different thresholds of $p(r|w)$ are shown in Figure 3. As can be seen, the MLE baselines (in red with syntactic patterns and green with intertext) perform consistently worse than the models learned using the topic models (in pink and blue). The difference in precision, aggregated across all relations, is statistically significant at 95% confidence for most of the thresholds.

Extractions aggregation We can take advantage of redundancy on the web to calculate a support metric for the extractions. In this experiment, for every extracted relation (r, e_1, e_2), for every occurrence of a pattern w_i connecting e_1 and e_2, we add up $p(r|w_i)$. Extractions that are obtained many times and from high-precision patterns will rank higher.

Table 1 describes the results of this aggregation. We have considered the top four highest-frequency relations for people. After aggregating all the extracted relations and ranking them by support, we have divided the evaluation set into two parts: (a) for relations that were not already in FreeBase, we evaluate the precision; (b) for extractions that were already in FreeBase, we take the top-confidence sentence identified and evaluate whether the sentence is providing support to the relation. For each of these, both syntactic patterns and intermediate-text patterns have been evaluated.

The results are very interesting: using syntax, *Death place* appears easy to extract new relations and to find support. The patterns obtained are quite unambiguous, e.g.



Relation	**Unknown relations**		**Known relations**	
	Correct relation P@50		Sentence support P@50	
	Syntax	*Intertext*	*Syntax*	*Intertext*
Parent	0.58	0.38	1.00	1.00
Death place	0.90	0.68	0.98	0.94
Birth place	0.38	0.56	0.54	0.98
Nationality	0.86	0.78	0.34	0.40

Table 1: Evaluation on aggregated extractions.

On the other hand, *birth place* and *nationality* have very different results for new relation acquisition vs. finding sentence support for new relations. The reason is that these relations are very correlated to other relations that we did not have in our training set. In the case of *birth place*, many relations refer to having an official position in the city, such as *mayor*; and for *nationality*, many of the patterns extract *presidents* or *ministers*. Not having *mayor* or *president* in our initial collection (see Figure 1), the support for these patterns is incorrectly learned. In the case of nationality, however, even though the extracted sentences do not support the relation (P@50 = 0.34 for intertext), the new relations extracted are mostly correct (P@50 = 0.86) as most presidents and ministers in the real world have the nationality of the country where they govern.

4 Conclusions

We have described a new distant supervision model with which to learn patterns for relation extraction with no manual intervention. Results are promising, we could obtain new relations that are not in FreeBase with a high precision for some relation types. It is also useful to extract support sentences for known relations. More work is needed in understanding which relations are compatible or overlapping and which ones can partially imply each other (such as *president-country* or *born_in-mayor*).

References

ACE. 2004. The automatic content extraction projects. http://projects.ldc.upenn.edu/ace.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI'07*.

D.T. Bollegala, Y. Matsuo, and M. Ishizuka. 2010. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In *Proceedings of the 19th international conference on World wide web*, pages 151–160. ACM.

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6, pages 9–16.

S. Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL*, volume 2007, pages 708–716.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The automatic content extraction (ace) program–tasks, data, and evaluation. In *Proceedings of LREC*, volume 4, pages 837–840. Citeseer.

A. Fader, S. Soderland, and O. Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of Empirical Methods in Natural Language Processing*.

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.

R. Hoffmann, C. Zhang, and D.S. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 286–295. Association for Computational Linguistics.

R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D.S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

G. Kasneci, M. Ramanath, F. Suchanek, and G. Weikum. 2009. The yago-naga approach to knowledge discovery. *ACM SIGMOD Record*, 37(4):41–47.

Z. Kozareva and E. Hovy. 2010. Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1482–1491. Association for Computational Linguistics.

C. Kwok, O. Etzioni, and D.S. Weld. 2001. Scaling question answering to the web. *ACM Transactions on Information Systems (TOIS)*, 19(3):242–262.

D. Li, S. Somasundaran, and A. Chakraborty. 2011. A combination of topic models with max-margin learning for relation detection.

P. McNamee and H.T. Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC)*.

D. Milne and I.H. Witten. 2008. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

J. Nivre. 2006. Inductive dependency parsing. In *Text, Speech and Language Technology*, volume 34. Springer Verlag.

M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. 2006. Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In *Proceedings of the National Conference on Artificial Intelligence*, page 1400. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

F. Pereira, A. Rajaraman, S. Sarawagi, W. Tunstall-Pedoe, G. Weikum, and A. Halevy. 2009. Answering web questions using structured data: dream or reality? *Proceedings of the VLDB Endowment*, 2(2):1646–1646.

H. Poon and P. Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 1–10. Association for Computational Linguistics.

D. Ravichandran and E. Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 41–47. Association for Computational Linguistics.

S. Sekine. 2006. On-demand information extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 731–738. Association for Computational Linguistics.

Beth M. Sundheim and Nancy A. Chinchor. 1993. Survey of the message understanding conferences. In *HLT'93*.

I. Titov and A. Klementiev. 2011. A bayesian model for unsupervised semantic parsing. In *The 49th Annual Meeting of the Association for Computational Linguistics*.

C. Wang, J. Fan, A. Kalyanpur, and D. Gondek. 2011. Relation extraction with relation topics. In *Proceedings of Empirical Methods in Natural Language Processing*.

Daniel S. Weld, Fei Wu, Eytan Adar, Saleema Amershi, James Fogarty, Raphael Hoffmann, Kayur Patel, and Michael Skinner. 2008. Intelligence in wikipedia. In *Proceedings of the 23rd national conference on Artificial intelligence*, pages 1609–1614. AAAI Press.

F. Wu and D.S. Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics.

L. Yao, A. Haghighi, S. Riedel, and A. McCallum. 2011. Structured relation discovery using generative models. In *Empirical Methods in Natural Language Processing (EMNLP)*.

A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland. 2007. Textrunner: Open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 25–26. Association for Computational Linguistics.