Freepal: A Large Collection of Deep Lexico-Syntactic Patterns for Relation Extraction

Johannes Kirschnick, Alan Akbik, Holmer Hemsen
Technische Universität Berlin
Database Systems and Information Management Group
Einsteinufer 17, 10587 Berlin, Germany
{firstname.lastname}@tu-berlin.de

Abstract

The increasing availability and maturity of both scalable computing architectures and deep syntactic parsers is opening up new possibilities for Relation Extraction (RE) on large corpora of natural language text. In this paper, we present FREEPAL, a resource designed to assist with the creation of relation extractors for more than 5,000 relations defined in the FREEBASE knowledge base (KB). The resource consists of over 10 million distinct lexico-syntactic patterns extracted from dependency trees, each of which is assigned to one or more FREEBASE relations with different confidence strengths. We generate the resource by executing a large-scale distant supervision approach on the CLUEWEB09 corpus to extract and parse over 260 million sentences labeled with FREEBASE entities and relations. We make FREEPAL freely available to the research community, and present a web demonstrator to the dataset, accessible from free-pal.appspot.com.

Keywords: Relation Extraction, Distant Supervision, Web Mining, Language resource

1. Introduction

We are currently witnessing three trends that point to increased potential for executing Relation Extraction (RE) on web scale text collections: Firstly, scalable computing architectures capable of processing ever larger amounts of data are being developed and becoming available (Dean and Ghemawat, 2004). Secondly, deep syntactic parsers are becoming more accurate and more robust, enabling their use on corpora of different domains, see Petrov and McDonald (2012) for a detailed discussion. And thirdly, ever larger datasets of raw web data on which RE can be performed are becoming readily available, such as CLUEWEB09 (Callan et al., 2009), SPINN3R (Burton et al., 2011) and COMMONCRAWL[1].

However, a major bottleneck to this potential is the effort involved in creating high quality relation extractors. In particular, for each relation of interest, a set of lexico-syntactic *patterns* must be identified that reliably indicates the presence of a relation instance in a sentence. Approaches for identifying such patterns range from manual rule-writing (Chiticariu et al., 2010) to learning extractors from text (Culotta and Sorensen, 2004; Min et al., 2013) using labeled training examples which are generated either manually or semi-automatically (Xu, 2008). In all these cases, the process is effort- and resource-intensive and must be repeated for every relation. This is especially of interest given the number of relations that can be defined, which can be as diverse and manifold as the text domains themselves; ranging from generic relations, such as PERSONSIBLINGOFPERSON or PARENTOF relationships, to very specific ones tailored to a domain of interest such as the FILMRELEASEDONMEDIUM relation.

Resource. In this paper, we present FREEPAL, a resource designed to assist with the creation of extractors for more than 5000 relations defined in the FREEBASE knowledge base (Bollacker et al., 2008). FREEPAL consists of over 10 million distinct lexico-syntactic patterns defined over dependency trees, each of which is assigned to one or more FREEBASE relations with different confidence strengths (see Table 1 for examples). We generate FREEPAL by executing a large-scale distant supervision approach (Mintz et al., 2009) on CLUEWEB09, a corpus of over 500 million web pages, using the FREEBASE knowledge base.

Contribution. Our intent is threefold: Firstly, by releasing this resource we aim to help research groups and individuals in tapping into the potential offered by RE on large corpora. Secondly, by showcasing our lexico-syntactic patterns in a publicly available web demonstrator we aim to engage the research community in a discussion on a suitable abstraction layer and feature set for defining RE patterns. Thirdly, we investigate the challenges and the potential of executing distant supervision for thousands of relations on large-scale data.

In the following, we outline our large-scale distant supervision approach, the challenges encountered with processing datasets at this scale and discuss the results.

2. Approach

We follow a three step distant supervision approach: First, we leverage FREEBASE to automatically annotate sentences with entity mentions and their relations to be used as training data. We then perform a pattern extraction step over all annotated sentences to gather lexico-syntactic patterns for all pairs of entities. Finally, we determine for each pattern the distribution over all FREEBASE relations it was observed with. This allows us to determine weighted pattern-relation assignments.

2.1. Generating Training Data

In order to find reliable lexico-syntactic patterns for FREEBASE relations, we seek to find as many sentences as possible that contain mentions of at least two FREEBASE entities. We make use of the recently released FACC1 (Gabrilovich et al., 2013) resource, a high quality named entity linking

[1] http://commoncrawl.org/

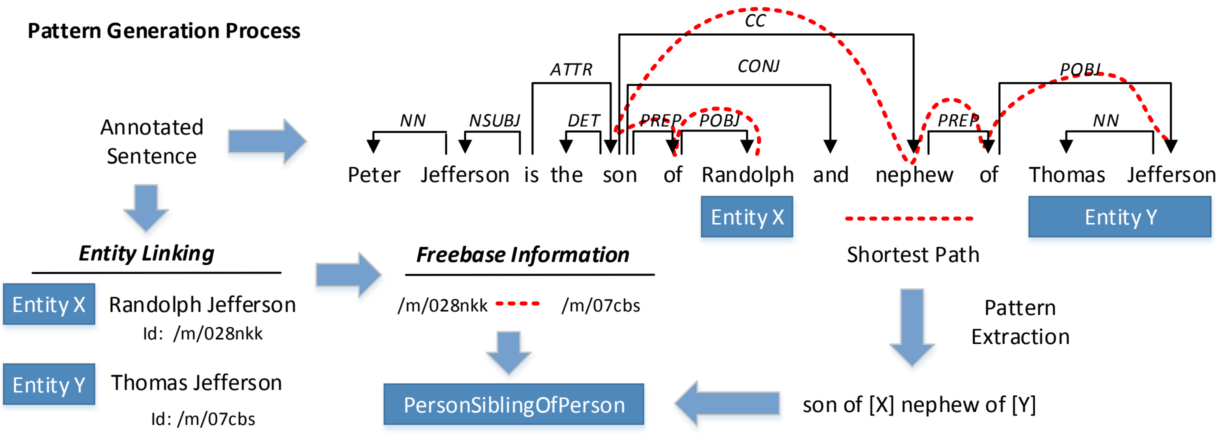


Figure 1: Patterns are extracted from the shortest undirected path between two annotated entities. The pattern is a candidate for all FREEBASE relations (in this case *PersonSiblingOfPerson*) that the two entities participate in.

effort that was executed on the CLUEWEB09 corpus, linking over 5 billion entity mentions to their corresponding FREEBASE entries.

Filtering Spam. Because a portion of CLUEWEB's 500 million web pages is spam, we employ the Waterloo spam ranking (Cormack, 2007) to identify and filter out such pages, while this filters quite aggressively, it prevents from processing text, which might have been automatically generated and thus does not provide new patterns varieties. On the remaining 135 million pages, we apply boilerplating (Kohlschütter et al., 2010) to remove all HTML markup and navigational elements, to extract the natural language text. This is followed by sentence splitting and a filter steps that removes all sentences that do not contain at least two FREEBASE entities.

Parsing. We use the CLEARNLP toolkit (Choi and McCallum, 2013) to perform dependency parsing on the remaining sentences, which uses a very fast parser with high accuracy. These steps produce a dataset of over 260 million parsed and annotated sentences using more than 5700 CPU compute hours.

2.2. Freebase Relationship Naming Convention

FREEBASE defines a rich set of typed binary relations that are identified by a concrete identifier. In the context of this paper, we show the abbreviated human readable identifiers for each relation, while the dataset contains the full one.

For example the detailed relation *people.person.sibling_s..people.sibling_relationship.sibling* represents the relation *people.person.sibling_s*, indicating that entity X is connected to entity Y via the sibling relationship, or informally Y is a sibling of X. More specifically it represent the subtype *people.sibling_relationship.sibling* of that relation. Subtypes arise, as a relationship can be further defined, for example by a start or end date, indicating when this relation holds. Here its the actual sibling that is being linked. To generate the abbreviated name, simply the main type is used and a sensible pattern is applied, yielding PERSONSIBLINGOFPERSON.

Throughout this paper we denote the governor of a relationship as Y and the dependent as X, or using a graphical view, X as left and Y as right participants of a relation.

2.3. Extracting Patterns

For each annotated and parsed sentence, we perform a pattern extraction step for all pairs of entities. We use a method that we previously employed in (Akbik et al., 2012; Akbik et al., 2013) in which we traverse the shortest undirected path in the dependency tree between two entities (Bunescu and Mooney, 2005). On the shortest path, we collect all lemmatized tokens and typed dependencies. The starting and ending positions in the shortest path are substituted by the entity placeholders [X] and [Y] respectively. The dependency trees capture short- as well as long-range entity dependencies present in the sentence while not mandating a certain word order.

Example. An example of the pattern extraction process is illustrated in Figure 1. Here, the input sentence *"Peter Jefferson is the son of* ***Randolph*** *and nephew of* ***Thomas Jefferson****"*, is annotated with the FREEBASE entities *Randolph* and *Thomas Jefferson* from the *FACC1* corpus. The shortest path between these entities is found by traversing the undirected path yielding the following dependency tree:

cc(son-1, nephew-4),
prep(son-1, of-2),
pobj(of-2, [X]-3),
prep(nephew-4, of-5),
pobj(of-5, [Y]-6)

Collecting the lemmatized tokens which are part of the path in left to right order generates the final pattern **SON OF [X] NEPHEW OF [Y]** and the associated dependencies.

We explicitly try not to generify the patterns any further, to define subsumption hierarchies or find inclusions. This is due to the fact that we are mapping patterns directly to FREEBASE relations, which themselves can be very broad or very fine grained, such as the CONTAINEDIN and

Pattern	Relation	Confidence
play [X] in movie [Y]	StarringInFilm	0.431
[X] title [Y]	GameDevelopedBy	0.299
watch [X] on [Y]	ProgramOnTVNetwork	0.204
[X] release of [Y]	FilmReleasedOnMedium	0.413
[X] be [Y] team	SportsTeamParticipatedInLeague	0.274
[X] name after [Y]	Namesakes	0.288
[X] subsidiary of [Y]	OrganizationChildOf	0.387
[X] receive nomination for [Y]	AwardNominationsFor	0.248
[X] [Y] pilot	MemberOfMilitaryForce	0.187
[X] administer by [Y]	ProtectedSitesGoverningBody	0.224
[X] division in [Y]	OrganizationHeadquarterTown	0.321
[X] protagonist of [Y]	BookCharacterIn	0.181
[X] marry in [Y]	MarriageLocation	0.171
perform [X] in [Y]	CharacterInOpera	0.218
son of [X] nephew of [Y]	PersonSiblingOfPerson	0.300

Table 1: Lemmatized forms of the top fifteen most common patterns. Only the lexical part of the patterns are displayed for readability reasons. The relation is the human readable form of the Freebase relation with the highest confidence.

SiteProtectedByGoverningBody relations respectively. Normalizing a pattern could lead to misclassification as it would clash with an already specified Freebase relation that we have found patterns for as well.

Using this method, we find over 10 million distinct patterns with their associated dependency tree that are observed at least three times in the corpus.

2.4. Assigning Patterns to Relations

In order to assign observed patterns to relations, we look up the Freebase relations of the entity pairs they were observed with[2]. We extract for each identified pair all relations. As relations are directed, we search for relations between X-Y and Y-X simultaneously. In some cases, two entities can participate in multiple relations; an example are the relations PersonNominatedForPrize and PersonWinsPrize that both often hold for the same entity pair. In such cases, we assign observed patterns to all possible relations. This processes inverse relationships as well, but only when they are expressed in Freebase. By repeating this procedure for all patterns, we determine for each pattern a distribution over Freebase relations.

Computing Pattern Assignments. We use this distribution as basis for computing the individual probabilities of a pattern-relation assignment as well as the overall entropy of a given pattern. So, for the given example, the son of [X] nephew of [Y] pattern is observed with the relations PersonSiblingOfPerson three times and GovernmentPositionAppointedByPerson once. Using the maximum likelihood estimation PersonSiblingOfPerson is chosen as the representative relation. This estimate serves as a confidence measure to indicate how plausible an assignment relation to pattern is.

2.5. Entropy Calculation

As the training data is skewed, just relying on the confidence measure for assignment alone is subject to misrepresentations. We therefore also calculate the information entropy over the observed relations for each pattern. Giving higher entropy for patterns that are observed with many relations and a lower one for patterns representing only a few ones, it thus measures how good a pattern can be explained by a set of relations.

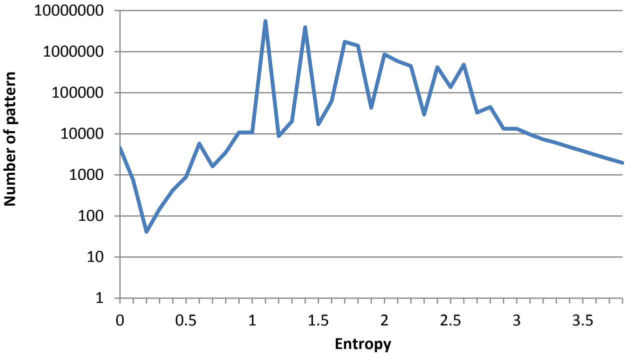


Figure 2: Histogram of the entropy distribution for the extracted pattern in logarithmic scale.

The entropy H for each pattern is calculated using the likelihood estimates P for each observed relation Rel_i in the following way:

$$H(Pattern) = -\sum_i P(Rel_i) \log P(Rel_i)$$

We experimentally determined 3.0 as a good cut-off point to filter non-descriptive from descriptive patterns, which were observed with too many different relations. The distribution of the entropy is shown in Figure 2. The histograms shows, that only a few pattern have a very small entropy, which is the case when the pattern has been observed only a few times or holds only for a very small number of different pattern. Removing patterns with a high entropy increases the quality of the resource while not removing too many patterns.

Our generated training data is skewed in the sense that we observe a power law distribution of relations in the corpus. This means that there are a few relations (such as Loca-

[2] We use a snapshot of Freebase retrieved on 05.05.2013

tionLocatedInLocation or PersonBornInLocation) that dominate Freebase, while other relations are more scarce. The pattern generation process effectively samples from this distribution for each pattern. We explicitly do not account for this as the entropy is already a good indicator for the entailed relation.

This becomes apparent for patterns which are seen with different relations, such as the pattern *[X] mayor of [Y]*. It is not uncommon that the mayor of a city was also born there and since many more entity pairs in Freebase express the PersonBornInLocation relation than the PersonMayorOfCity, we see the same skew.

Results. Filtering by entropy gives us for each pattern the top relations it points to and their associated confidence. Table 1 lists the pattern as well as the identified target relation for the fifteen most frequently observed pattern-dependency pairs in the corpus.

Our process generates high quality, descriptive pattern for most of the well-defined relations. As the resource is too large to completely evaluate, this is supported by manual inspection of the most common relations.

3. Discussion

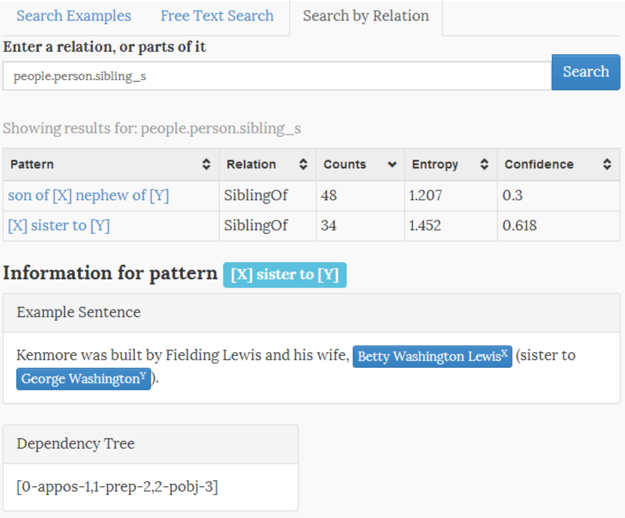
We made a number of observations when applying distant supervision at scale to determine learning lexico-syntactic patterns for thousands of relations. In the following, we highlight the two most important observations.

Doubly skewed data. We found the occurrence of different relation types to be heavily skewed; prominent relations are observed very frequently, giving us a very good basis for determining patterns, while many other relations are not. Similarly, the occurrence of patterns is skewed, with some patterns being observed frequently for a given relation, while many others are rarer. One effect of this double skew is that only 68,621 lexico-syntactic patterns were seen more than 50 times. Considering the very large amount of data we conducted this effort on (i.e. over 5 billion entity mentions in more than 130 million web pages), we had expected a higher number. This observation points to difficulties, possibly even limitations, of the use of distant supervision to identify less common lexico-syntactic patterns in the long tail of all possible patterns that point to a relation.

Effort and cost. Of all data preprocessing steps, we found the dependency parsing of the 260 million labeled sentences to be the most costly in terms of CPU hours. In terms of implementation effort, we found the linking of the ClueWeb and FACC1 resources, as well as the extraction of sentences that contain at least two entities, to be the most challenging. By contrast, the actual pattern extraction and computation of pattern-relation assignments is relatively straight forward and processes in about 3 CPU hours. This means that having this infrastructure in place allows us to quickly experiment with different extraction strategies.

4. Demonstration and Outlook

With our web demonstrator, available at `free-pal.appspot.com`, we showcase the patterns found using distant supervision at scale. For each pattern an example sentence can be selected to understand the source of the pattern as well as example participating entities. Figure 3 shows a snapshot of the UI. A large subset comprising of 23.000 distinct pattern can be interactively queried, filtered for individual tokens or relations and sorted on entropy or confidence.



Enter a relation, or parts of it

people.person.sibling_s

Showing results for: people.person.sibling_s

Pattern	Relation	Counts	Entropy	Confidence
son of [X] nephew of [Y]	SiblingOf	48	1.207	0.3
[X] sister to [Y]	SiblingOf	34	1.452	0.618

Information for pattern [X] sister to [Y]

Example Sentence

Kenmore was built by Fielding Lewis and his wife, Betty Washington Lewis^X (sister to George Washington^Y).

Dependency Tree

[0-appos-1,1-prep-2,2-pobj-3]

Figure 3: Web demonstrator, to interactively query and filter a subset of the derived pattern.

The dataset is released to the research community to evaluate our method further as well as use the findings as basis for building powerful relation extraction systems. We see a straight forward application of the resource in identifying relations in unlabeled corpora using named entity recognizers. Furthermore, as each relation is typed within Freebase this can be exploited to perform entity disambiguation by ruling out impossible entity types based on the expected relationship types.

For future work, we see two main avenues of research: The first is to experiment with different ways of defining lexico-syntactic patterns for RE, possibly with the help of feedback by the user community through our demonstrator. Presently, we are investigating to model selectional restrictions (SR) into patterns similar to (Akbik et al., 2013), as well as defining patterns that incorporate inter document references in addition to sentence-level dependencies. The second is to develop methods that allow the community to use our patterns to effortlessly create relation extractors that can immediately be deployed to large text corpora.

5. Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. Johannes Kirschnick and Holmer Hemsen received funding from the German Federal Ministry of Economics and Technology (BMWi) under grant agreement A01MD11018, ’A cloud-based Marketplace for Information and Analytics on the German Web’ (MIA). Alan Akbik received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement ICT-2009-4-1 270137 ’Scalable Preservation Environments’ (SCAPE).

6. References

Akbik, A., Visengeriyeva, L., Herger, P., Hemsen, H., and Löser, A. (2012). Unsupervised Discovery of Relations and Discriminative Extraction Patterns. In *Proceedings of the 24th International Conference on Computational Linguistics*.

Akbik, A., Visengeriyeva, L., Kirschnick, J., and Löser, A. (2013). Effective Selectional Restrictions for Unsupervised Relation Extraction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Bunescu, R. C. and Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731. Association for Computational Linguistics.

Burton, K., Kasch, N., and Soboroff, I. (2011). The ICWSM 2011 Spinn3r dataset. In *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2011)*.

Callan, J., Hoy, M., Yoo, C., and Zhao, L. (2009). The ClueWeb09 Dataset. Available online at `http://lemurproject.org/clueweb09/`.

Chiticariu, L., Krishnamurthy, R., Li, Y., Raghavan, S., Reiss, F. R., and Vaithyanathan, S. (2010). SystemT: an algebraic approach to declarative information extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 128–137, Stroudsburg, PA, USA. Association for Computational Linguistics.

Choi, J. D. and McCallum, A. (2013). Transition-based dependency parsing with selectional branching. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria*.

Cormack, G. V. (2007). University of Waterloo Participation in the TREC 2007 Spam Track. In *Proceedings of the 16th Text REtrieval Conference*. Dataset available online at `https://plg.uwaterloo.ca/~gvcormac/clueweb09spam/`.

Culotta, A. and Sorensen, J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics.

Dean, J. and Ghemawat, S. (2004). MapReduce: simplified data processing on large clusters. In *Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04, pages 137–150, Berkeley, CA, USA. USENIX Association.

Gabrilovich, E., Ringgaard, M., and Subramanya, A. (2013). FACC1: freebase annotation of ClueWeb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0).

Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450. ACM.

Min, B., Grishman, R., Wan, L., Wang, C., and Gondek, D. (2013). Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. In *Proceedings of NAACL-HLT*, pages 777–782.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (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.

Petrov, S. and McDonald, R. (2012). Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, volume 59.

Xu, F.-Y. (2008). *Bootstrapping Relation Extraction from Semantic Seeds*. Ph.D. thesis, Saarland University.