Learning a Theory of Marriage (and other relations) from a Web Corpus

Sandro Bauer[1], Stephen Clark[1], Laura Rimell[1], Thore Graepel[2]

[1] University of Cambridge, Cambridge, United Kingdom
firstname.lastname@cl.cam.ac.uk
[2] Microsoft Research Cambridge, Cambridge, United Kingdom
thore.graepel@microsoft.com

Abstract. This paper describes a method for learning which relations are highly associated with a given seed relation such as *marriage* or *working for a company*. Relation instances taken from a large knowledge base are used as seeds for obtaining candidate sentences expressing the associated relations. Relations of interest are identified by parsing the sentences and extracting dependency graph fragments, which are then ranked to determine which of them are most closely associated with the seed relation. We call the sets of associated relations *relation theories*. The quality of the induced theories is evaluated using human judgements.

1 Introduction

Information Retrieval, and related areas such as Information Extraction and Question Answering, are starting to move away from "shallow" approaches based on keywords to more semantically informed approaches. One example is the use of entailment rules, allowing "Tom Cruise divorced Katie Holmes", for example, to be recognized as providing the answer to the question "Who did Tom Cruise marry?" (through the rule "X divorces Y" implies "X was married to Y"). Taking this example further, suppose the text says that "X has given birth to her fifth child with Y". Whilst not entailing that X is married to Y, there is some likelihood that X and Y are indeed married. Knowledge such as this constitutes what we are calling a *relation theory*. Inducing such a theory is in line with the more general vision of Machine Reading [1], which aims to allow language processing systems to make many of the inferences that humans make when processing text.

The aim of our work is to infer these tacit associations automatically from text, with a score for each associated relation indicating the strength of the association. More concretely, it is a set of such (*relation, score*) pairs for the marriage relation that we call a *theory of marriage* in this paper. The associated relations are in the form of dependency graph fragments. Our method is inspired by the *distant supervision hypothesis* [2], which assumes that all sentences containing instances of a relation do express that relation (e.g. that all sentences containing *William* and *Kate* express the fact that Prince William and Kate Middleton are

married). However, rather than maintaining this hypothesis, we argue that its apparent *failure* in many contexts can be used to our advantage, because many inaccurate examples are semantically close to the seed relations. For example, a sentence stating that William gave Kate an engagement ring can be used as evidence that *giving an engagement ring* is associated with the marriage relation.

2 Methodology

Our experiments exploited three freely-available resources. Freebase[3], a crowd-sourced knowledge base, was the source of entity pairs standing in the seed relation. ClueWeb09[4], a corpus of 500 million English web pages, provided text from which candidates for the associated relations were extracted. Finally, a large background corpus of parsed sentences from Wikipedia was used to rank candidate relations. We will use the marriage relation as a running example.

First, the ClueWeb09 corpus was processed using the boilerplate removal tool in [3], together with some additional simple pre-processing steps such as removing overly long or short sentences. Next, all ClueWeb sentences were extracted which contained references to any pair of entities standing in the marriage relation in Freebase. This resulted in a set of 1,022,271 sentences, which provided the source from which to extract the additional relations associated with marriage.

In terms of how to extract the additional relations, we considered two alternatives. First, we experimented with Open IE tools, in particular ReVerb [4] and OLLIE [5], an extension to ReVerb based on dependency paths. These tools often correctly detected associated relations, but missed some examples, particularly those involving long-range dependencies. Also, the flat output in the form of triples makes it difficult to generalise over syntactically similar relations (e.g. "be girlfriend of" and "be long-time girlfriend of"). We therefore decided to build our own extraction tool based on full dependency graphs. Our representation, like that of OLLIE, is an extension to the notion of dependency path in [6].

First, all the extracted sentences from ClueWeb were parsed with the C&C parser [7], which produces typed dependency graphs. Then all sentences with no dependency path between the two entities were removed. Using a parser in this way we obtain around twice as many extractions as with ReVerb. For the remaining sentences, the direct dependency path between the two married entities is extracted, and a number of heuristics are used to add side nodes from the dependency path, i.e. nodes connected to one of the nodes on the dependency path through a single edge. Specifically, we add common-noun direct objects of verbs (where an indirect object or the second object of a ditransitive verb is on the path) and prepositions with no further outlinks. This way, we capture phrasal verbs such as "X puts up with Y" (which would otherwise be reduced to "X puts with Y") and verbal constructions such as "X ties knot with Y", where the direct object "knot" is crucial for the meaning of the relation.

[3] `http://www.freebase.com/`
[4] `http://lemurproject.org/clueweb09/`

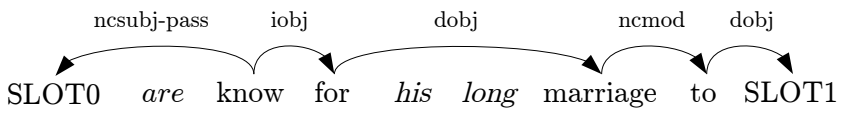

Fig. 1: Example of a dependency graph fragment, with lemmas as tokens

This process results in dependency graph fragments such as the one shown in Figure 1 (from the sentence *Mr Sauer is known for his long marriage to Merkel*): The tokens making up the dependency graph fragment are linked through typed dependency edges, and the SLOTs represent the married entities from the KB.

In order to induce the tacit associations, we employ association metrics typically used for finding collocations [8]. The idea is that we want to rank highly those graph fragments which appear more often in the marriage corpus (the set of sentences containing instances of married entities) than would be expected in a general corpus. We used a parsed version of Wikipedia as the background corpus, and calculated the number of times that the relevant graph fragments occur in this corpus. We tried several association measures, and found that the standard t-test produced promising results. As a final filter, graph fragments which occured less than 5 times in the corpus, or which appeared with less than 5 entity pairs, were removed. Note that the induced relations are not canonical, in the sense that more than one graph fragment can denote the same relation, and we leave clustering of the fragments for future work.

3 Human Evaluation

In order to verify that the ranked lists of dependency graph fragments capture relations that humans associate with the given seed relations, we asked annotators to judge the quality of the induced rankings. Thirteen annotators, all computational linguists, participated, and each annotator was asked to evaluate a total of 60 graph fragments taken from two different seed relations. Participants were given two example sentences per fragment, with the entities and the words in the graph fragment marked with different colours. The task was to assess whether the relation represented by the fragment is highly, somewhat or not at all associated with the seed relation. There was a fourth option for cases where there is an associated relation in at least one of the sentences, but our heuristics failed to capture all the words representing it (see Table 1 for examples).

Our method was evaluated on four Freebase relations: *marriage*, *parent*, *employment by a company*, and *birthplace*. For each seed relation, we evaluated the 100 most highly ranked graph fragments and the first 10 of every 100 fragments outside the top 100 ("less highly ranked fragments"). The total number of fragments varied from 429 to 1,084 depending on the seed relation. Each annotator was presented with, for each of two seed relations, 20 graph fragments from the top 100 and 10 fragments from the less highly ranked fragments, in random order. A subset of the data (10 of the 39 chunks) was presented to a second set of annotators to measure inter-annotator agreement. We calculated percentage agreement per rank, averaged across all relations, as well as the κ coefficient [9].

Table 1: Example sentences for the marriage relation with annotator ratings. Words in dependency graph fragment are in bold; entities from KB are in italics.

rating	example
highly	*Sonia Gandhi* **is** the **widow of** *Rajiv Gandhi* who was assassinated in 1991.
somewhat	In October 2007 *Miranda Kerr* **started dating** English hottie *Orlando Bloom*.
not at all	*Nicoletta Braschi* **worked with** *Roberto Benigni* in a lot of his films.
wrong words	UsMagazine.com is reporting that *Amy Winehouse* **did** marry boyfriend *Blake Fielder-Civil* in Florida yesterday.

Table 2: Examples from the top 100 graph fragments for the four relations. Words in dependency graph fragments are in bold; entities from KB are in italics.

relation	examples
marriage	Baywatch hottie *Pamela Anderson* **tied** a **knot with** fiance *Kid Rock* this weekend. *Lauryn Hill* **gave birth** over the weekend **to** her fifth **child with** *Rohan Marley*.
parenthood	As God had commanded, *Abraham* **circumcised** *Isaac* when he was eight days old. ... and *Liev Schreiber* have **welcomed** their second son, *Samuel Kai Schreiber*.
birthplace	*Harold Washington* (1922-1987) **was** the first African-American **mayor of** *Chicago*. *Han Hoogerbrugge* is a digital artist **living in** *Rotterdam*, Netherlands.
employment	*Buzz Aldrin* **retired from** *NASA* as long ago as 1972. *Scott Gnau* **is** vice president and general **manager** ... **at** *Teradata*.

Tables 3a to 3d give the percentages at each rank for the four relations (e.g. 42% of the top 100 graph fragments for the marriage relation were judged as highly relevant). For all four relations, more than half of the top 100 fragments in the lists were judged highly or somewhat associated. Table 2 gives some examples from the top 100 graph fragments for each of the relations. In contrast, the less highly ranked fragments are overwhelmingly rated as somewhat associated or below. Some relations show an especially quick drop-off: the *marriage* relation has many examples such as *write letter to* and *save* in the less highly rated associated fragments, which could apply to a married couple, but are arguably not part of a core theory of marriage.

Figure 2 shows the precision at rank summed over all four relations, where a point is awarded when a fragment is judged highly or somewhat associated. For the agreement experiment, we obtained an overall percentage agreement score of 67.5% and a κ coefficient of 0.55, indicating a fair level of agreement.

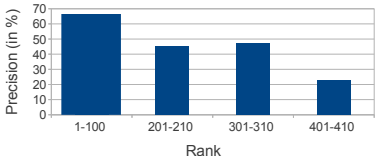

Fig. 2: Precision at rank across all relations

4 Related Work

There is a large body of broadly related work attempting to induce general entailment rules, such as *X writes Y* implies *X is the author of Y* [6]; or if a company is based in a city, and the city is located in a state, then the company is headquarted in that state [10, 11]. Whilst our work can be situated in this broad area of IE, it is more closely associated with attempts to derive "domain theories" from text [12], which might contain rules stating that, for example, flights do not start and end in the same city; or attempts to induce rules between relations

Table 3: Scores for the different relations

rank	highly	somewhat	not at all	wrong words
1-100	42	29	23	6
201-210	0	60	30	10
301-310	20	10	40	30
401-410	0	10	90	0
501-510	0	10	60	30

(a) *marriage*

rank	highly	somewhat	not at all	wrong words
1-100	19	33	42	6
201-210	10	20	50	20
301-310	30	40	30	0
401-410	10	10	70	10
501-510	10	20	70	0
601-610	10	40	40	10
701-710	0	0	90	10
801-810	0	20	80	0
901-910	0	10	90	0
1001-1010	0	20	80	0

(b) *birthplace*

rank	highly	somewhat	not at all	wrong words
1-100	43	8	31	18
201-210	10	0	70	20
301-310	0	10	80	10
401-410	0	10	50	40

(c) *parenthood*

rank	highly	somewhat	not at all	wrong words
1-100	70	22	2	6
201-210	40	40	10	10
301-310	40	40	10	10
401-410	30	20	40	10

(d) *employment*

in a KB, such as: if two people have children in common, then they are often married [13]. Chambers et al.[14] exploit co-reference in documents to extract narrative schemas in an unsupervised setting, for example the events associated with a criminal being arrested. In some ways the failure of the distant supervision hypothesis could also be seen as important for their system, since it relies on the semantic association of sentences involving the same actors; however, in our system we use seed instances from a KB in order to test whether richer representations of relations can be bootstrapped from a KB.

5 Conclusion

We have described a novel method for inducing theories of relations that enhance the information present in KBs, which we believe is a valuable ingredient for more semantically informed IR systems. The theories presented here are simplified versions of what an ideal relation theory would contain. Obvious extensions include making the theories culture- and era-dependent; for example, divorce is much more associated with marriage in the modern era than in the past. Similarly, we could include temporal ordering; for example the fact that engagement happens before and divorce occurs after and is the end of marriage.

As a step in this direction, we performed a small pilot study making use of the birthdates of people available in Freebase. We induced theories of marriage for different time periods, including the biblical era and the 20th century, by seeding the extraction with only couples born in the relevant era. The results reflect how marriage depends on historical context. Relations such as *commit sin with* or *bury one's partner* are more frequent for protagonists in the bible, and are ranked highly for the biblical theory, while *suing one's partner* and *being spotted with one's partner* are more highly ranked for the contemporary theory.

Acknowledgments. The first author was supported by Microsoft Research through its PhD Scholarship Programme.

References

1. Etzioni, O., Banko, M., Cafarella, M.J.: Machine Reading. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence. AAAI'06, Boston, Massachusetts, USA (2006) 1517–1519
2. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: 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. ACL '09, Suntec, Singapore (2009) 1003–1011
3. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate Detection using Shallow Text Features. In: Proceedings of the third ACM international conference on Web search and data mining. WSDM '10, New York, New York, USA (2010) 441–450
4. Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11, Edinburgh, United Kingdom (2011) 1535–1545
5. Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open Language Learning for Information Extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. EMNLP-CoNLL '12, Jeju Island, Korea (2012) 523–534
6. Lin, D., Pantel, P.: DIRT – Discovery of Inference Rules from Text. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001, Location unknown (2001) 323–328
7. Clark, S., Curran, J.R.: Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. Computational Linguistics **33**(4) (2007) 493–547
8. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, Massachusetts, USA (1999)
9. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20** (1960) 37–46
10. Schoenmackers, S., Etzioni, O., Weld, D.S., Davis, J.: Learning first-order horn clauses from web text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. EMNLP '10, Cambridge, Massachusetts, USA (2010) 1088–1098
11. Berant, J., Dagan, I., Goldberger, J.: Global learning of typed entailment rules. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Portland, Oregon, USA (2011) 610–619
12. Liakata, M., Pulman, S.: Learning theories from text. In: Proceedings of the 20th international conference on Computational Linguistics. COLING '04, Geneva, Switzerland (2004)
13. Galárraga, L., Teflioudi, C., Suchanek, F., Hose, K.: AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases. In: Proceedings of the 22nd International World Wide Web Conference (WWW 2013), Rio do Janeiro, Brazil (2013)
14. Chambers, N., Jurafsky, D.: Unsupervised Learning of Narrative Schemas and their Participants. 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. ACL '09, Suntec, Singapore (2009) 602–610