model, with $p \ll 0.05$. We also note that despite using the same assumption, the joint model performs much worse than the distant supervision approach in this scenario. Learning a model of relations and mentions is inherently more difficult. Using a wrong assumption will hence more likely hurt performance.

Does the at-least-once model help to fix the type of error discussed in section 2? To find out, we inspect the results of the \textit{founded} relation. When we consider the top 100 instances of this relation for the distant supervision system, we observe a precision of 45%. Compare this to 72% precision for the at-least-once model.

On close inspection, most of the distant supervision errors for the \textit{founded} relation stem from cases where patterns such as “director of” appear. They indicate that the person in question works for the given company. Because in the training set such patterns often appear when a person is a founder, they gain high weights and appear high up in the ranking.

The at-least-once model also makes this type of error, but to a much lesser extent. This is not surprising if we consider that for training instances with only one mention, the at-least-once and distant supervision assumptions are equivalent. Assume that according to Freebase, person A founded company B. If there is only one mention of A and B in the NYT training corpus, it has to be a mention of \textit{founded}, even if the sentence says “director-of”. This leads to a higher weight for “director-of” as \textit{founded} pattern.

7 Conclusion

This paper presents a novel approach to extract relations from text without explicit training annotation. Recent approaches assume that every sentence that mentions two related entities expresses the corresponding relation. Motivated
by the observation that this assumptions frequently does not hold, in particular when considering external knowledge bases, we propose to relax it. Instead we assume that at least one sentence which mentions two related entities expresses the corresponding relation.

To model this assumption we make two contributions. First, we introduce a novel undirected graphical model that captures both the task of predicting relations between entities, and the task of predicting which sentences express these relations. Second, we propose to train this graphical model by framing distant supervision as an instance of constraint-driven semi-supervision. In particular, we use SampleRank, a discriminative learning algorithm for large factor graphs, and inject the expressed-at-least-once assumption through a truth function.

Empirically this approach improves precision substantially. For the task of extracting 1000 Freebase relation instances from the New York Times, we measure a precision of 91% for at-least-once supervision, and 87% for distant supervision. This amounts to an error reduction rate of 31%.

A crucial aspect of our approach is its extensibility: framed exclusively in terms of factor graphs and truth functions, it is conceptually easy to apply it to larger tasks such as the joint prediction of relations and entity types. In future work we will exploit this aspect and extend our model to jointly perform other relevant tasks for the automatic construction of KBs.

Acknowledgements. This work was supported in part by the Center for Intelligent Information Retrieval, in part by SRI International subcontract #27-001338 and ARFL prime contract #FA8750-09-C-0181, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and in part by UPenn NSF medium IIS-0803847. Any opinions, findings and conclusions or recommendations expressed in this material are the authors’ and do not necessarily reflect those of the sponsor.

References