source is external. The New York Times data contains over 1.8 million articles written and published by the New York Times between January 1, 1987 and June 19, 2007. Generally, we find that Freebase entities are frequently mentioned in the NYT corpus. For example, for the year 2007 about 700,000 mentions of Freebase entities appear in the corpus. Naturally, we observe a smaller number of cases in which two related entities are mentioned in the same sentence: again for the year 2007 we find about 170,000 such cases.

6.2 Preprocessing

In order to find entity mentions in text we first used the Stanford named entity recognizer [12]. The NER tagger segments each document into sentences and classifies each token into four categories: PERSON, ORGANIZATION, LOCATION and NONE. We treat consecutive tokens which share the same category as single entity mention. Then we associate these mentions with Freebase entities. This is achieved by simply performing a string match between entity mention phrase and the canonical names of entities in Freebase.

Next, for each pair of entities participating in a relation of our training KB, we traverse the text corpus and find sentences in which the two entities co-occur. Each pair of entity mentions is considered to be a relation mention candidate. For each such candidate we extract a set of features (see section 3). The types of features are essentially corresponding to the ones used by [18]: we used lexical, Part-Of-Speech (POS), named entity and syntactic features (i.e. features obtained from the dependency parsing tree of a sentence). We applied the openNLP POS tagger4 to obtain POS tags and used the MaltParser [20] for dependency parsing.

6.3 Held-Out Evaluation

Following [18] we divide the Freebase relations into two parts, one for training and one for testing. The former is aligned to the years 2005-2006 of the NYT corpus, the latter to the year 2007. As candidate relation instances we use all pairs of Freebase entities that are at least once mentioned in the same sentence. Note that the amount of Freebase relations mentioned in the training set (4700) and test set (1950) is relatively low due to a smaller overlap between Freebase and the New York Times. Hence we cannot evaluate our models with the same quantity of data as [18].

In figure 2 we compare the precision and recall curve for the baseline distant-supervision model (distant), the supervised joint model (joint) and the distant model with expressed-at-least-once assumption (at-least-once). The curve is constructed by ranking the predicted relation instances using their loglinear score. For the distant supervision baseline this score is first normalized by the number of mentions.5 We traverse this list from high score to low score, and measure precision and recall at each position.

---

4 Available at http://opennlp.sourceforge.net/

5 This yielded the best results for the baseline. We also tried to use conditional probabilities to rank. This lead to poor results because SampleRank training has no probabilistic interpretation.
We can see that the model with expressed-at-least-once assumption is consistently outperforming the distant supervision baseline and the supervised joint model. This suggests that the at-least-once model has the best sense of how relations that are already contained in Freebase are expressed in NYT data. However, it does not necessarily mean that it knows best how relations are expressed that are not yet in Freebase. We address this in the next section.

6.4 Manual Evaluation

For manual evaluation all Freebase entities and relations are used as training instances. As candidate relation instances we choose those entity pairs which appear together in the NYT test set, but for which least one participating entity is not in Freebase. This means that there is no overlap between the held-out and manual candidates. Then we apply our models to this test set, and asked two annotators to evaluate the top 1000 predicted relation instances.

We cannot calculate recall in this case, since we cannot provide all relation instances expressed in our corpus. Instead we use a “Precision at K” metric with respect to the ranked lists we extracted in section 6.3. Figure 3 shows the precisions for values of K between 0 and 1000.

We first note that the precision is much higher for manual evaluation than for held-out evaluation. This shows that false negatives in Freebase are an issue when doing held-out evaluation. Many of the false positives we predict are in fact true relation instances and just do not appear in Freebase.

For manual evaluation the at-least-once model is still the winner. At K = 1000 we observe a precision of 91% for at-least-once supervision, 87% for distant supervision. This amounts to an error reduction rate of 31%. The sign test shows that the at-least-once model is significantly better than the distant supervision