Crosslingual Distant Supervision for Extracting Relations of Different Complexity

Andre Blessing
Institute for
Natural Language Processing
Universität Stuttgart
Pfaffenwaldring 5b
Stuttgart, Germany
blessing@ims.uni-stuttgart.de

Hinrich Schütze
Institute for
Natural Language Processing
Universität Stuttgart
Pfaffenwaldring 5b
Stuttgart, Germany
supervision@ifnlp.org

ABSTRACT

We propose *crosslingual distant supervision* (crosslingual DS) for relation extraction, an approach that automatically extracts labels from a pivot language for labeling one or more target languages. The approach has two benefits compared to standard DS: (i) increased coverage if target language labels are not available; and (ii) higher accuracy of automatically generated labels because noisy labels are eliminated in crosslingual filtering. An evaluation for two relations of different complexity shows that crosslingual DS increases the accuracy of relation extraction. Our approach is language independent; we successfully apply it to four different languages: Chinese, English, French and German.

Categories and Subject Descriptors

I.2.6 [**ARTIFICIAL INTELLIGENCE**]: Learning—*knowledge acquisition*; I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing

Keywords

crosslingual distant supervision, relation extraction

1. INTRODUCTION

Most relation extraction systems are statistical and need to be trained on a training set that represents all required relations. For most relations, such training sets are not available; this is particularly true of languages other than English. Unsupervised learning [4] is not a general solution to this problem as unsupervised relation extractors often have low accuracy or fail to discover the required relations. In this paper, we use *distant supervision* (DS) [34, 23] to address this problem. Our innovation is that we apply it to a *crosslingual setting* – in contrast to prior monolingual work. We show that crosslingual DS increases accuracy and coverage of DS.

DS uses existing knowledge bases (KBs) to label free text and then train a supervised model on the labels. Figure 1 shows an example of standard (monolingual) DS. An instance of the relation – <Paris, capital-of, France> – is retrieved from the KB. Next, unstructured data referring to the arguments of the relation is identified in raw text and the text is annotated (shown as markup) with the instance from the KB. We call this annotation an automatically generated label or *automatic label*. A corpus annotated in this way is then used to train a relation classifier.

Our innovation in this paper is to perform DS crosslingually: a KB in a *pivot language* is used to annotate text in a *target language*. The approach has two benefits compared to standard DS: *higher accuracy* and *increased coverage*.

Crosslingual DS achieves higher accuracy of automatic labels due to crosslingual filtering or *CL-filtering* that occurs during label transfer from the pivot to the target. A training example is generated only if pivot language KB, translation and target language raw text are consistent with respect to that training instance. This makes it unlikely that erroneous training examples are generated. We will present extensive experimental evidence for the benefit of CL-filtering in our evaluation below.

Turning to the second point, crosslingual DS increases coverage in cases where the target language KB has no coverage of a relation or less than the pivot language KB. Differences in coverage are widespread because KBs in any given language tend to lack instances of the relation that are less relevant to the region or culture where the language is spoken – e.g., Islais Creek in San Francisco has no French Wikipedia article and the Parisian rivulet Grange-Batelière has no English Wikipedia article. There also exist differences in the coverage of KBs that are due to the community-based and sometimes random nature of what is modeled; e.g., there is no comprehensive German KB for the spouse-of relation.

Because of these differences in coverage of relations in KBs and Wikipedias across languages, crosslingual methods have a great potential of increasing coverage. If there is no KB for a relation in a particular language, then monolingual DS cannot be applied. By extending monolingual DS crosslingually, the advantages of DS (automatic generation of large training sets for supervised machine learning) become available to languages that currently do not have good KBs for many or most relations.

Evidence that there is a need to increase coverage of the standard monolingual approach is given in [18]. They showed in the context of the 2010 Knowledge Base Population shared task (KBP2010) that Wiki-derived databases cover only well-known entities. For example, Freebase covers only 48% of the slot types and 5% of the slot answers in the KBP2010 evaluation data. Crosslingual DS addresses this problem by increasing coverage.

The two benefits of crosslingual DS are of particular importance for what we call *complex relations*, which are harder to learn and

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

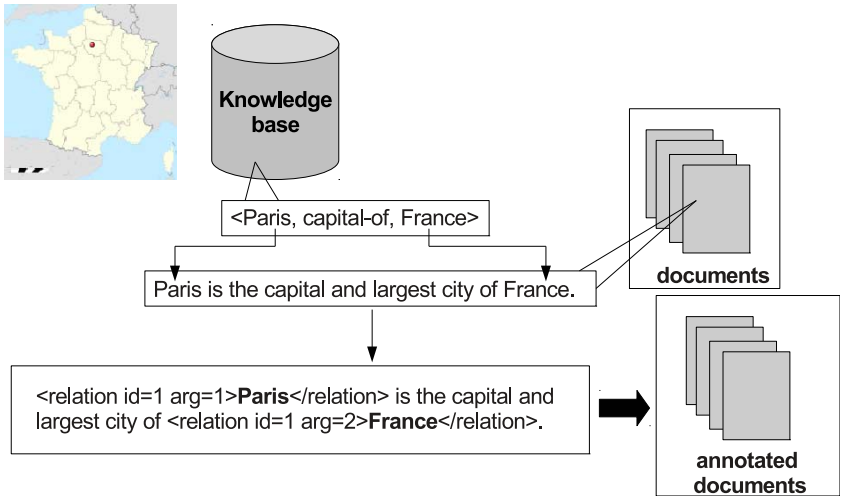

Figure 1: Automatic annotation in distant supervision

extract than the simple relations that most work on relation extraction has focused on. In this paper, we investigate spouse-of 'person X is spouse of person Y' as a representative of a simple relation; and river-town 'settlement X is on river Y' as a representative of a complex relation. There is no simple categorical property that distinguishes simple and complex relations. Rather, there are a number of dimensions along which relations vary; and river-town is more complex on each of those dimensions than spouse-of.

One dimension is that marriage imposes a strict constraint on the categories of its arguments: only human beings can be married. River-town is more complicated in that there are a number of related, but distinct relations between geopolitical entities and bodies of water; e.g., *the Rhine flows through the Netherlands* or *Cleveland on Lake Erie*. When recognizing an instance of river-town, the extraction system must verify the correctness of the categories of the two arguments. Ontological variability of arguments also causes higher variability in the linguistic means used to express the relation: different words will be used to express that a village is located on the Nile vs that Islais Creek is in San Francisco.

Another important dimension is that marriage is a central fact of life. If a person is the main subject of a Wikipedia article, then it is likely that their spouse(s) will be described as well. This is less likely for river-town: many Wikipedia articles about rivers do not contain an exhaustive list of the towns on that river. This means that spouse-of is more predictable and has more training material for DS than river-town.

Prominence of a relation and ontological simplicity have two further consequences that distinguish simple and complex relations. First, simple relations are more likely to have a consistent and formally correct modeling in Wikipedia infoboxes. This is the case for spouse-of in the English Wikipedia. In contrast, river-town is modeled in a number of different ways in infoboxes: attribute names are inconsistent and attribute values are often difficult to parse – e.g., if a value is a list of strings that are not hyperlinked.

Second, freely available KBs provide high-quality data for DS for simple relations, but often do not for complex relations. Again, this is the case for spouse-of (which has good coverage in Freebase) and river-town (which is not systematically covered in any existing KB). Rigorousness of infoboxes and availability in KBs are of course highly correlated because KBs are to a large extent based on infoboxes.

Most work on DS and unsupervised relation extraction has been done on simpler relations like spouse-of. We will take advantage of this prior work and use an existing gold standard for spouse-of. In contrast, little work has been done on complex relations like river-town. One of the contributions of this paper is that we systematically conduct experiments for both types of relations. We will show below that the benefits of crosslingual DS (in particular, better coverage and higher accuracy) are of particular value for complex relations. Complex relations have fewer instances that can be automatically labeled using only the target language; hence they benefit from the additional automatic labels provided by the pivot language. Complex relations also suffer more from incorrect automatic labels than simple relations because they are ontologically more complex. Crosslingual DS counteracts this effect by filtering out incorrect labels.

In the next section, we introduce DS in more detail. Section 3 describes the three steps of crosslingual DS. Experiments on a simple and a complex relation are described in Section 4 and evaluated in Section 5. Sections 6 and 7 discuss related work and state our conclusions.

2. DISTANT SUPERVISION (DS)

DS uses existing KBs to label free text and then train a supervised model on the labels. A KB commonly used for this purpose is Freebase [8], but any reliable source of structured data (e.g., Wikipedia infoboxes) can be used. Closely related to DS are self-training [21, 19] and self-annotation [10].

There are two different DS approaches. In what we call *redundancy-based DS* [23], all matches of a KB fact in the text are annotated – without applying any filtering criteria. This results in noisy labels and requires the use of robust learning algorithms that can learn accurate classifiers from noisy data if the training set is large and redundant enough. Redundancy-based DS works well for simple relations as long as an appropriate KB in the target language exists; but it is unclear how it can be applied to complex relations and to simple relations without KB coverage in the target language.

An alternative approach is what we call *structure-based DS* [31,

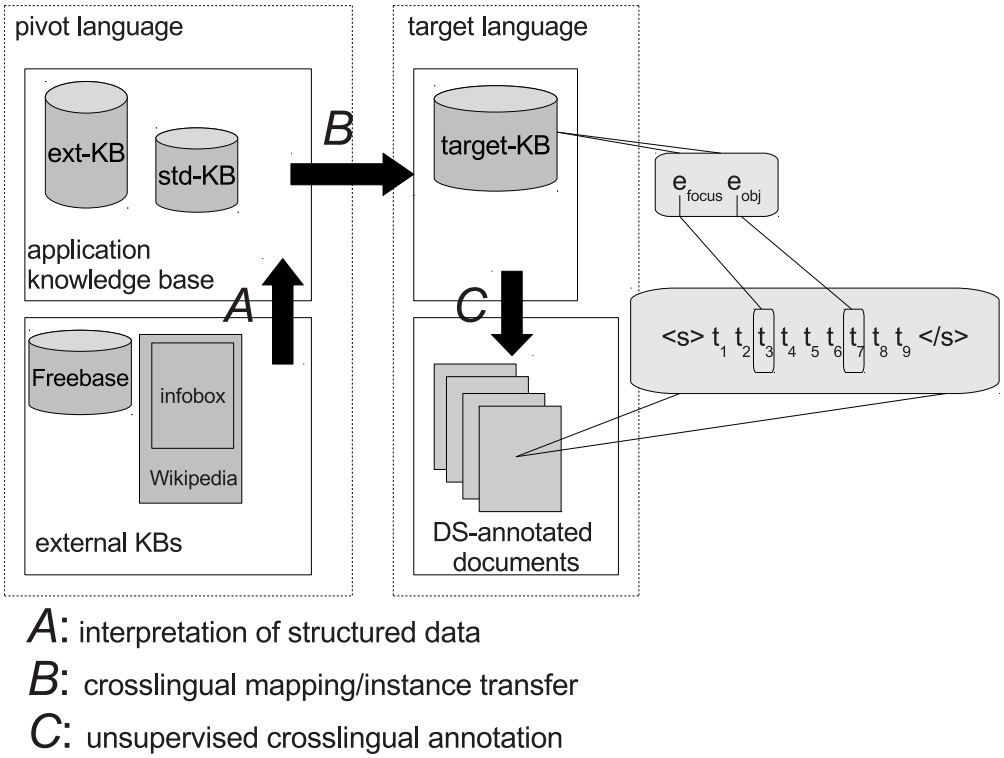

Figure 2: The three steps of crosslingual DS (A, B, C), discussed in detail in Sections 3.1–3.3

39]. It exploits additional structural knowledge in Wikipedia to eliminate noisy labels. Specifically, it requires that there be a dependency – a hyperlink or any other structural linkage – between the KB fact and the textual resource. With such structural constraints, each individual annotation can be made with high confidence and the resulting annotated training set is highly accurate. We adopt structure-based DS as our approach in this paper.

A disadvantage of structure-based DS compared to redundancy-based DS is that it has lower coverage because many annotations that do not meet structural constraints are filtered out. Crosslingual DS addresses this problem by crosslingual transfer of instances. In addition, we address this problem in this paper by developing "high-yield" methods for extracting KB facts for complex relations from Wikipedia infoboxes.

Wikipedia infoboxes are special kinds of templates that represent facts and relations of the article's entity in attribute-value pairs. Wiki-derived KBs like DBpedia [3], Yago [35] and Freebase are to a large extent based on relations extracted from infoboxes. These KBs are mostly used for redundancy-based DS (e.g., [23]). We will show in Section 5 that more extensive mining of the contents of infoboxes yields good results – especially for complex relations – that cannot be achieved using facts from Wiki-derived KBs.

3. CROSSLINGUAL DS METHOD

Figure 2 shows the three steps of crosslingual DS. After selecting a pivot language, we extract instances of the requested relation from an appropriate KB (Section 3.1, A in Figure 2); the KB can be a publicly accessible KB like Freebase, Wikipedia infoboxes or another (semi)structured data source. Depending on the quality of the extracted instances, this data is stored in the two different application specific knowledge bases standard(std)-KB and extended(ext)-KB, to be described in Section 3.1. In the transfer step (Section 3.2, B), a crosslingual mapping is used to translate the instances and transfer them to target-KB. In the final step (Section 3.3, C), the instances in target-KB are used to label relations on the sentence level in the target language (right half of figure, labeling step is magnified: "$e_{\text{focus}} e_{\text{obj}}$" and "$< s > t_1 \ldots t_9 < /s >$").

3.1 Interpretation of structured data

We first need to select a pivot language for which a KB with good coverage of the relation exists. Infoboxes in Wikipedia are often a good choice because Wikipedia's more than 200 language editions contain a wide range of fine-grained knowledge for many regions and cultures. English, French, German and Spanish are particularly well suited as pivot languages because their Wikipedia editions define a wide variety of infoboxes that are instantiated in a large number of articles. Of the languages that provide an infobox of the right type, we select the language with the highest coverage. This is then the source of arrow A in Figure 2.

Once the pivot language KB has been identified, we populate the application KBs std-KB and ext-KB. std-KB is used for clean data with clear semantics. There are two types of KBs that have this property: (i) high-quality databases like Freebase, DBpedia and Yago and (ii) Wikipedia infoboxes that are consistently and cleanly modeled. We define consistent and clean modeling as (i) there is a single consistent attribute name for the relation used in the infobox and (ii) the value of the attribute is a single hyperlinked entity or a list of hyperlinked entities.

Instances of complex relations that are not cleanly modeled are stored in a separate database ext-KB. Complex relations are usually not covered in public KBs. Since they are not cleanly modeled in infoboxes, automatic extraction (which produces near error-free re-

attribute	value
Großstädte 'large cities'	nächstgelegene Großstadt: [[Reutlingen]] 'nearest large city:' [[Reutlingen]]

Figure 3: Example of a "nonrigorous" attribute value in a Wikipedia infobox. See text for discussion.

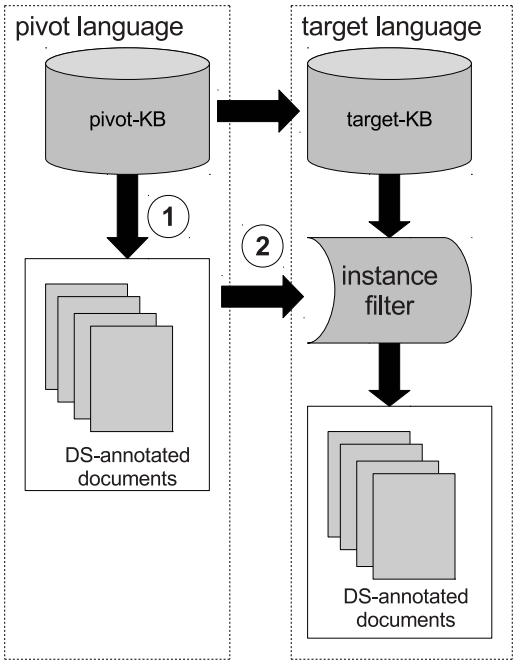

Figure 4: CL-filtering step: An instance (pair of entities) passes through the instance filter and is then used for annotation only if both pivot and target articles contain it.

sults for simple relations like spouse-of) is error prone. One reason they are not cleanly modeled is schema drift: no consistent naming scheme for the attributes of the relation is enforced. Another is that attribute values often include free text. This is a barrier to automatic extraction because the words in the free text are often ambiguous names or are modified by negation or other linguistic means. As an example consider the river infobox in the German Wikipedia, which describes up to 50 different attributes of rivers. Of the 2074 infoboxes with values for the attribute "Großstädte" 'large cities' (which lists all large cities on the river) 1030 contain exactly one hyperlink, 778 contain enumerations of hyperlinks (separated by delimiters that vary from infobox to infobox), and the remaining 264 contain all kinds of different formats: names without hyperlinks, further descriptions, notes that no such city exists etc.

Figure 3 is an example of a nonrigorous infobox: an entry of the river infobox in the German Wikipedia article about the river Erms. For this river, there is no appropriate value for the attribute 'large cities'. In such cases, editors sometimes put related information that is not compatible with the semantics of the attribute. In this case, they put "Reutlingen", a nearby city. As a consequence, the semantics of the attribute are no longer clear and consistent.

To populate ext-KB, we first create an inventory of all hyperlink anchor names (the *anchor inventory*) for rivers and towns in Wikipedia (cf. [37]); this inventory covers possible surface representations used for the entities. The infobox extractor then extracts all phrases that match an entry in the anchor inventory. Each such anchor phrase found in the value of settlement attributes (for river articles) and river attributes (for settlement articles) is stored as one instance of river-town in ext-KB. This infobox extraction increases coverage compared to std-KB, but generates noisy instances in cases like Erms-Reutlingen.

As described in detail below, we prevent the nonrigor of complex-relation infoboxes from decreasing precision by crosslingual filtering or *CL-filtering*: we filter out pivot language labels that cannot be consistently transferred and applied to the target.

3.2 Instance transfer

After populating std-KB and ext-KB, the pivot language instances are mapped to target-KB (B in Figure 2). We do this using Interwiki links following a number of recent papers that show how to deal with uncertainty and asymmetry of links (e.g., [13], [29]). Since the Wikipedias are not parallel corpora, many instances cannot be transferred from pivot to target; e.g., no appropriate Interwiki links may exist because one of the two entities does not have an article in the target.

Using Wikipedia for mapping is the easiest solution for this step. If Wikipedia is not sufficient, transliteration methods from statistical machine translation can be used (e.g., [16]).

3.3 Unsupervised crosslingual annotation

We refer to the entity a Wikipedia article is about as the *focus entity* or e_{focus} and to the second argument of the relation as e_{obj}; e.g., if "Paris is located on the Seine" occurred in the Wikipedia article about Paris: e_{focus}=Paris, e_{obj}=Seine.

The instances in target-KB are used to annotate relations in the unstructured text of the target language Wikipedia (C in Fig. 2). An effective heuristic for high-precision automatic annotation is to only consider sentences that contain e_{focus} – considering other sentences increases recall, but at a high cost to precision. A second filter that increases precision is that we only consider mentions of an e_{obj} that are hyperlinked to e_{obj}'s Wikipedia article.

e_{focus} is – in contrast to e_{obj} – usually not represented as a hyperlink. Recognizing e_{focus} is difficult because the title often has add-ons that disambiguate the entity (e.g., "Tom Jackson (politician)"). We address this problem by using the anchor inventory introduced in Section 3.1: a phrase in the article is interpreted as e_{focus} if it matches an entry in the inventory that is hyperlinked to e_{focus}. Finally, we perform a simple form of coreference resolution: we assume that all pronouns that match the main entity in gender and number refer to e_{focus}. If the gender and number of the main entity is unknown, then the most frequent pronoun is used for the reference. We have found that this simple heuristic has a high accuracy of between 95% and 100%.

We have so far only discussed how to create positive labels: target-KB facts only give rise to positive labels. We use a simple heuristic for negative labels building on [14, 36]. First, we infer the entity's type from the entity's article's infobox – e.g., an entity is a river if its article contains a river infobox. This heuristic works well because infoboxes are widely used in Wikipedia. An instance is negative if the inferred type of e_{obj} violates the constraints of the relation. This is the case if e_{obj} is not a settlement (if e_{focus} is a river) or is not a river (if e_{focus} is a settlement) for river-town; and if it is not a person for spouse-of.

3.4 CL-filtering

We can now precisely state how CL-filtering works. A pair of entities is only used for annotation if both pivot and target contain a sentence that contains both e_{focus} and e_{obj}. In practice, we first

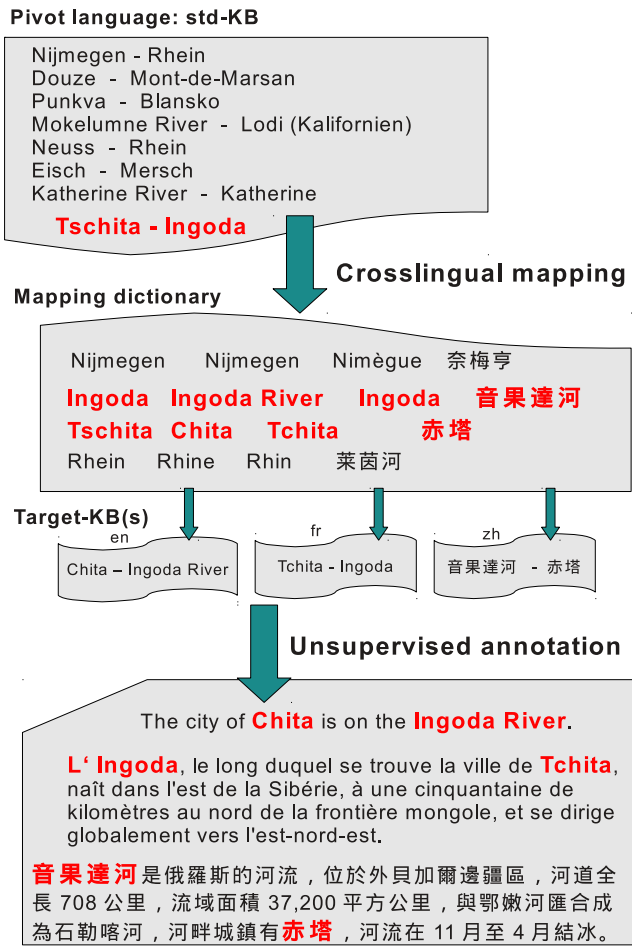

Figure 5: Illustrative example for crosslingual DS annotation for the river-town relation between the Indoga river and the town Chita. This instance is part of a German knowledge base. In the transfer step ("Crosslingual mapping"), the names are translated to English, French and Chinese by a mapping dictionary. In the "Unsupervised annotation" step, a sentence in each of the three target languages is automatically annotated.

annotate a corpus in the pivot language (left half of Figure 4). Then an additional filtering step is used during the annotation process on the target language (right half of Figure 4). For each possible annotation in the target language, we check if the corresponding instance (pair of entities) is also annotated in the pivot language. The annotation is rejected if there is no corresponding instance. As a result, the annotated training corpus in the target language contains only annotated instances that also occur in the pivot language.

Figure 3 is an example of the beneficial effect of CL-filtering. The river Erms and the city Reutlingen occur in the same sentence in the German Wikipedia and would be incorrectly labeled as a positive instance of river-town in a monolingual approach. But they do not occur in the same context in the English Wikipedia and CL-filtering therefore blocks the labeling. This is an example of CL-filtering applied to the *pivot*; see Table 1, de1–de3. The evaluation below will show that CL-filtering is also beneficial for the *target*.

3.5 Illustrative example

Figure 5 illustrates crosslingual DS as we have described it in Sections 3.1–3.3 for a concrete example. We consider the river-town relation between the Ingoda river and the city Chita. Since this relation is only modeled in the German Wikipedia, the only way to use distant supervision for a language other than German is a crosslingual approach. Since the relation occurs in the German Wikipidea, German is the pivot language.

Our goal is to annotate sentences in three target languages: English, French and Chinese. Since we have already given detailed examples for the knowledge base population step in Section 3.2, we start our walk-through with a pivot language KB (in this case: std-KB) in the upper part of Figure 5 that is already populated. All terms (e.g., "Nijmegen", "Lodi (Kalifornien)") refer to German Wikipedia articles. In the instance transfer step B (see Section 3.2, "Crosslingual mapping" in the figure), we have to find the matching articles for each of the target Wikipedias. The center part of Figure 5 depicts some entries of the mapping dictionary. We used only Interwiki links to generate this dictionary, but other transliteration methods would also be appropriate. The populated target KBs after instance transfer are shown below the mapping dictionary.

In the bottom part of Figure 5 the corresponding text passages in the articles are annotated by the method described in Section 3.3. As a result, we end up in each of our three target languages with a sentence annotated by crosslingual DS. These annotated sentences can then be used to train a native relation extractor.

Algorithm 1 Crosslingual DS

for all $r_{\text{pivot}}(e_i,e_j) \in \text{KB}_{\text{pivot}}$ **do**
  $\text{KB}^{+}_{\text{target}} \mathrel{+}= \{\text{map}(r_{\text{pivot}}(e_i,e_j))\}$
end for
for all $s \in \text{CORPUS}_{\text{target}}$ **do**
  for all $(e_{\text{focus}},e_{\text{obj}}) \in s$ **do**
    if $r_{\text{target}}(e_{\text{focus}},e_{\text{obj}}) \in \text{KB}^{+}_{\text{target}}$ **then**
      $\text{ANNOTATED-CORPUS}^{+}_{\text{target}} \mathrel{+}= \{< s,e_{\text{focus}},e_{\text{obj}}>\}$
    else if wrong-type(e_{obj}) **then**
      $\text{ANNOTATED-CORPUS}^{-}_{\text{target}} \mathrel{+}= \{< s,e_{\text{focus}},e_{\text{obj}}>\}$
    end if
  end for
end for

Algorithm 1 gives pseudocode for crosslingual DS. The first loop is used to map all entity pairs from the pivot language KB to the target language KB using the methods of Section 3.2. The next loop is used for the annotation process described in Section 3.3. We iterate over all sentences of the target corpus. In each sentence, all possible pairs of e_{focus} and e_{obj} are considered. If target-KB contains a pair, then a positive annotation is created. If the pair is not a member of target-KB and if the type of e_{obj} is not admissible for the coresponding slot of the relation, then a negative annotation is created.

Algorithm 2 gives pseudocode for the version of algorithm 1 that includes CL-filtering. We first collect all instances in the pivot language corpus, map them to the target and then store them in filtered target-KBs for positive instances ($\text{KB}^{+}_{\text{target-filtered}}$) and negative instances ($\text{KB}^{-}_{\text{target-filtered}}$). The second part of the algorithm is similar to the second part of Algorithm 1: a pair $(e_{\text{focus}},e_{\text{obj}})$ of entities is used for a positive annotation only if $\text{KB}^{+}_{\text{target-filtered}}$ contains it and for a negative annotation only if $\text{KB}^{-}_{\text{target-filtered}}$ contains it. This ensures that each target annotation is justified by a target language sentence as well as a pivot language sentence.

Algorithm 2 Crosslingual DS extended by CL-filtering

for all $s \in \text{CORPUS}_{\text{pivot}}$ **do**
  for all $(e_{\text{focus}}, e_{\text{obj}}) \in s$ **do**
    if $r_{\text{pivot}}(e_{\text{focus}}, e_{\text{obj}}) \in \text{KB}_{\text{pivot}}$ **then**
      $\text{KB}^{+}_{\text{target-filtered}} \mathrel{+}= \{\text{map}(r_{\text{pivot}}(e_i, e_j))\}$
    else if wrong-type(e_{obj}) **then**
      $\text{KB}^{-}_{\text{target-filtered}} \mathrel{+}= \{\text{map}(r_{\text{pivot}}(e_i, e_j))\}$
    end if
  end for
end for
for all $s \in \text{CORPUS}_{\text{target}}$ **do**
  for all $(e_{\text{focus}}, e_{\text{obj}}) \in s$ **do**
    if $r_{\text{target}}(e_{\text{focus}}, e_{\text{obj}}) \in \text{KB}^{+}_{\text{target-filtered}}$ **then**
      $\text{ANNOTATED-CORPUS}^{+}_{\text{target}} \mathrel{+}= \{< s, e_{\text{focus}}, e_{\text{obj}} >\}$
    else if $r_{\text{target}}(e_{\text{focus}}, e_{\text{obj}}) \in \text{KB}^{-}_{\text{target-filtered}}$ **then**
      $\text{ANNOTATED-CORPUS}^{-}_{\text{target}} \mathrel{+}= \{< s, e_{\text{focus}}, e_{\text{obj}} >\}$
    end if
  end for
end for

4. EXPERIMENTS

We conduct experiments for the *complex* relation river-town and the *simple* relation spouse-of. While it is a limitation of this paper that we conduct experiments on only one relation of each type, it is important to consider that creating large test sets in four languages is a considerable effort.

4.1 River-town relation

River-town is the relation between a river and a settlement – a village, town or city – located on the river, e.g., (Seine, Paris). This relation is needed in applications like the semantic web and in geographic information systems (GIS). As a GIS use case for river-town consider the 2010 cholera outbreak in Haiti. The source of the outbreak was the Artibonite river; in a case like this an international aid organization will need to obtain a complete list of all the settlements on the river. Even though there are geo-spatial datasets such as OpenStreetMap available, we are not aware of any that represent river-town. The crosslingual DS method introduced in this paper addresses this problem.

For the experiments on river-town, we use the target languages Chinese, English and French. These languages have no (Chinese, English) or not enough (French) infobox data on river-town for monolingual DS to work. Freebase also does not represent this relation. In contrast, the German Wikipedia does represent river-town in infoboxes; thus, we can use it as pivot. To address the nonrigorousness of river-town infoboxes discussed above, we use CL-filtering to increase accuracy of automatic labeling in DS.

4.2 Spouse-of relation

As an example of a simple relation, we run experiments on the well-known spouse-of relation [12, 18]. In contrast to river-town, spouse-of is clear-cut and not open to interpretation in the cultures represented in Wikipedia. We ignore temporal aspects, so the actual relation we extract is "X was Y's spouse at any time in the past". The German Wikipedia has no structured information about spouse-of and does not include it in infoboxes. So we have to consult a pivot language to enable an automatic annotation of German by DS. In this case English is a good pivot since it provides enough spouse-of instances in Freebase that are linked to English Wikipedia articles.

4.3 Data

We use JWPL [40] to process Wikipedia[1] and extended it for recursively nested infoboxes. For river-town, we only consider Wikipedia articles about rivers and towns. The selection process is done in several stages. First we populate std-KB and ext-KB with instances from German infoboxes. These instances are split into three sets: training (80%), development (10%) and test (10%). The development set was used to develop the feature representation. To ensure a fair evaluation, we impose the constraint that any given entity only occur in one of training, development or test. Next we collect for each entity in target-KB the corresponding Wikipedia articles in the target and pivot languages. The resulting training set contains 1580 Chinese, 4068 English, 2673 French and 5741 German Wikipedia articles for river-town. CL-filtering mandates that a given instance occur in both pivot and target language articles. Thus, each of the non-German articles has a corresponding German article.

The test set contains 330 Chinese, 1299 English, 858 French and 1302 German articles. English, French and German test sets were annotated by two annotators ($.77 \leq \kappa \leq .88$); for Chinese only one annotator was available.

For spouse-of, we use the set of [12] as test set. It consists of 78 Wikipedia articles. We then extract 40,000 relation instances from Freebase[2] and remove all instances that occur in the 78 articles. The next steps are analogous to the river-town selection processing. A total of 3741 German and English articles survive CL-filtering and can be used as training set. To reduce the amount of manual annotation necessary for spouse-of, we translate German entity pairs found by the extractor in testing back to pairs of English entities using Interwiki links. A German pair extracted from the test set is defined to be correct if and only if its translation via Interwiki links is a member of the Culotta set.

4.4 Relation extraction

We use the maximum entropy classifier Mallet [22] for relation extraction. Each instance, consisting of a pair of named entities, is represented as a feature vector. Our feature representation is similar to that of [41]. It includes information about the tokens in a window around and between the two entities, including POS tags and capitalization. Additional features capture the dependency path between the two entities and all tokens on this path.

4.5 Computational Cost

To reduce implementation effort we use a previously developed DS system [5] that uses the UIMA [15] framework as the basis for this paper. The ClearTK [26] toolkit is our interface to Mallet [22].

The computationally most costly step of crosslingual DS is the linguistic processing of the textual data: sentence splitting, tokenizing, tagging and parsing using tools described in [32, 7]. It takes several hours. The linguistically annotated corpora are stored in a relational database which enables fast access and avoids the need to repeat any of the expensive linguistic processing. In comparison, the automatic annotation by DS is quite fast and takes only a few minutes. The training process and application of the model to the test data is also very fast and takes under a minute.

[1]We used dumps of the Chinese (20110412), English (20110405), French (20110409) and German (20110410) Wikipedias.

[2]We use Freebase to ensure comparability with earlier work. We confirmed that we can extract a very similar data set directly from Wikipedia using simple extraction.

5. EVALUATION

Our evaluation measures are

- precision: TP/(TP+FP)
- recall: TP/(TP+FN)
- F_1: harmonic mean of precision and recall

A pair that was correctly (resp. incorrectly) classified as being an instance of the relation is a true positive TP (resp. a false positive FP). An instance of the relation not recognized by the classifier is a false negative FN.

5.1 Results for river-town

Table 1 shows results for river-town. To compare automatic DS labels with manual labels, we manually annotated the training data in the pivot (German). These instances are transferred to the target languages and classifiers are trained on the resulting three annotated sets. This is referred to as "human" in the table.

There are three main results of the experiments. (i) Crosslingual DS using ext-KB produces classifiers that have performance comparable to or better than classifiers trained on labels transferred from manual labels in the pivot language: lines {zh,en,fr}1 vs {zh, en,fr}3 in Table 1. The largest decrease is for English unfiltered (unfil.) F_1 and recall (72.3 vs 70.8 and 61.2 vs 59.3); in all other cases the difference is smaller or crosslingual DS is clearly better. The fact that there is so little difference between manual pivot labels transferred to the target and automatic labels transferred to the target is evidence for the high quality of crosslingual DS labels.

(ii) The ext-KB approach (which exploits all infobox information, not just a clean and easily parsable subset) is superior in F_1 to the standard approach std-KB (which only uses clean infoboxes): lines {zh,en,fr}3 vs {zh,en,fr}2 in Table 1. Precision is slightly worse for English and French, but recall and F_1 are consistently better. Increases in F_1 are at least 5% and much larger in several cases. Chinese benefits the most because the number of transferable instances is much smaller than for the other two languages, resulting in a relatively small std-KB. This result suggests that the extended approach performs better than standard DS for complex relations like river-town – relations that are more likely to have nonrigorous infoboxes.

(iii) CL-filtering (fil.) further improves ext-KB F_1 by about 2% for French, 5% for English and 8% for Chinese (lines {zh,en,fr}3 in Table 1). Each filtered F_1 score that is significantly better than the corresponding unfiltered score at $p < .05$ is bold (approximate randomization test [25] as implemented in [27]).

The main effect of CL-filtering is that it increases the recall of the trained classifier because it eliminates many false negative training instances. Recall that a negative label is only applied if the two entities occur in the same sentence in both target and pivot Wikipedia articles. In the case of Chinese, CL-filtering increases both recall and precision because there is a larger number of false positives (similar to Erms-Reutlingen) that are eliminated.

		F_1		precision		recall	
		unfil.	fil.	unfil.	fil.	unfil.	fil.
zh1	human	40.0	**55.4**	50.0	62.1	33.3	50.0
zh2	std-KB	9.8	**34.8**	40.0	80.0	5.6	22.2
zh3	ext-KB	56.6	**64.3**	88.2	90.0	41.7	50.0
en1	human	72.3	**74.3**	88.5	86.7	61.2	65.1
en2	std-KB	61.0	**66.5**	89.0	87.8	46.4	53.6
en3	ext-KB	70.8	**75.5**	87.8	87.1	59.3	66.6
fr1	human	65.0	**70.2**	86.2	87.3	52.2	58.7
fr2	std-KB	60.4	**64.5**	89.1	86.6	45.6	51.4
fr3	ext-KB	67.1	69.8	88.6	87.5	54.0	58.1

Table 1: Performance of crosslingual DS on river-town for Chinese (zh1-zh3), English (en1-en3), and French (fr1-fr3).

		F_1		precision		recall	
		unfil.	fil.	unfil.	fil.	unfil.	fil.
de1	human	57.4	**66.7**	87.5	84.6	42.7	55.1
de2	std-KB	37.2	**46.1**	89.9	85.7	23.4	31.5
de3	ext-KB	57.5	**66.0**	84.6	80.9	43.6	55.8

Table 2: Performance of unfiltered vs filtered "monolingual" DS. Filtering in this case filters German annotations using the English data.

		F_1		precision		recall	
		unfil.	fil.	unfil.	fil.	unfil.	fil.
1	de	70.1	71.0	61.8	69.1	81.0	73.1
2	en (pivot)	57.5	**62.4**	45.9	55.7	77.8	70.8

Table 3: Performance of crosslingual DS on spouse-of

Table 2 shows that CL-filtering is even beneficial in a "monolingual" setting – that is, if it is applied to pivot language text: $F_1 =$ 57.5 (unfil) vs 66.0 (fil) on line de3.[3] Again, the reason for this effect is that many noisy German instances are removed by CL-filtering because they do not have English equivalents. This confirms the potential of using crosslingual information in DS.

The success of ext-KB (compared to std-KB and human) on the three language pairs shows that the approach can be applied to different languages with equal success and suggests that the basic design of our method is language-independent.

5.2 Results for spouse-of

For spouse-of, we perform experiments for the target language German and use English as the pivot. Table 3 shows that (i) crosslingual DS supports the creation of a well performing classifier on the target language for which no KB that contains spouse-of exists; (ii) CL-filtering is beneficial for the pivot language (significantly) as well as for the target language.

These results suggest that crosslingual DS is a promising approach for both simple relations like spouse-of and complex relations like river-town.

5.3 Impact of training set size

Figures 6–7 show that – except for very small training sets in English – filtered is consistently better than unfiltered in crosslingual DS ("filtered" curves vs "unfiltered" curves). The figures also demonstrate the high quality of the ext-KB labels: the ext-KB classifier is very close to (or better than) the classifier trained on human labels transferred to target-KB for training set sizes ≥ 350.

Figure 8 shows the performance of "monolingual" DS when pivot language annotations (German) are filtered using a target language

[3]To avoid confusion, we keep referring to German as the pivot and English as the target here even though their roles are partially reversed. We use the term "monolingual" to stress that one major benefit of crosslingual DS – positive instances are transferred from one language to the other – does not apply here. But even this monolingual variant of crosslingual DS is of course still crosslingual in the sense that it uses crosslingual filtering. We used English for CL-filtering here, but verified that F_1 is similar if French is used instead.

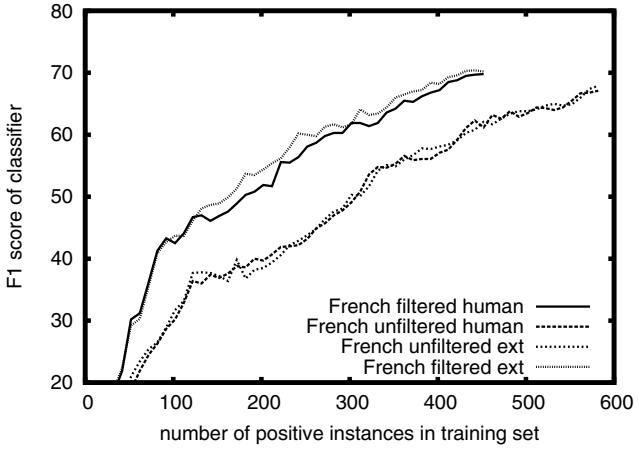

Figure 6: river-town: F_1 for French (target) as a function of training set size

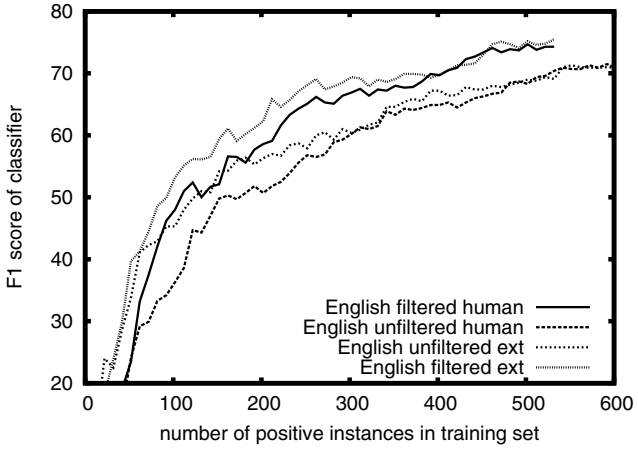

Figure 7: river-town: F_1 for English (target) as a function of training set size

(English). In this setting a pair of entities is only used for annotation in monolingual DS if there is a sentence in the target language that also contains both entities. In our case we use German as the language for the monolingual annotation and the English corpus for filtering. The difference to the unfiltered – purely monolingual – approach is clearly observable. This shows that crosslingual distant supervision can also help to improve the quality of monolingual distant supervision.

Finally, we show the performance for the spouse data (Figure 9). There is also an improvement by using filtering, but the impact is lower than for river-town. The reason is that in comparison with river-town more sentences represent possible instance candidates since there is a large number of possible relations between persons. The current filtering approach cannot distinguish instances on such a fine-grained level. This problem will be addressed in future work.

5.4 Comparison to SMT

We cannot directly compare crosslingual to monolingual DS since there is no appropriate structured data available for monolingual DS in the target; e.g., there is no river-town structured data for English.

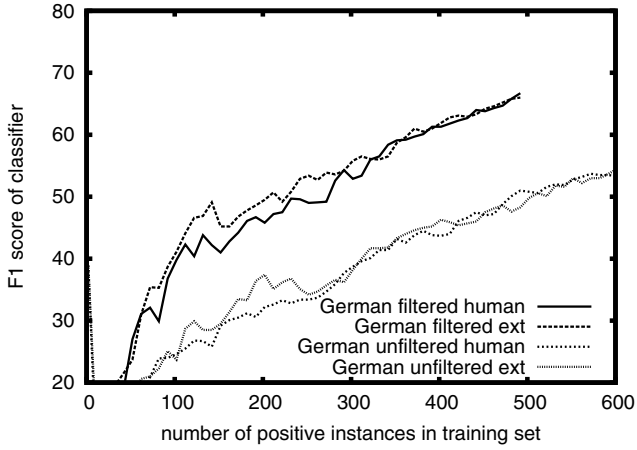

Figure 8: river-town: F_1 for German (pivot) as a function of training set size

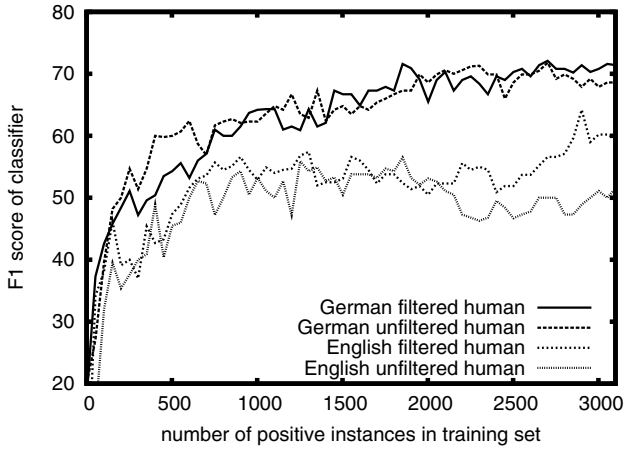

Figure 9: spouse-of: F_1 for English (target) and German (pivot) as a function of training set size.

However, we can automatically create a complete English monolingual setting for river-town using statistical machine translation (SMT) [28]. We used a subset of the test set, a set of 30 English Wikipedia articles, for evaluation. This set was translated into German using Google Translate. All e_{focus} and e_{obj} were manually marked in the translated sentences. Then the "human" model for German (Table 1, de1) was used as a classifier. In Table 4, we compare the result of this experiment ("SMT DS") with the classifier ext-KB (Table 1, en3), run on the test set of 30 articles. Crosslingual DS F_1 is about twice as large as SMT DS F_1.

If SMT produced perfect translations, we would expect performance comparable to crosslingual DS. The fact that SMT DS performs so much worse indicates that the quality of SMT is not yet good enough to be competitive with crosslingual DS.

6. RELATED WORK

A significant amount of research has been done on how to avoid the need for manually labeled training data for relation extraction.

	F_1	precision	recall
SMT DS	34.4	48.1	27.7
crosslingual DS	69.8	73.1	66.7

Table 4: SMT DS vs crosslingual DS for river-town

Two approaches to this problem are unsupervised methods [9, 4, 38] and distant supervision [23, 39, 17, 24]. In contrast to our approach, the prior work on DS is monolingual.

Craven and Kumlien [11] used automatic ("distant") labels for automatic annotation in bioinformatics. The term "distant supervision" was introduced by Mintz et al. [23]. They put DS on a more generally applicable footing by using a knowledge base (Freebase) that provides more encyclopedic facts and relations for automatic annotation. Other recent work that has contributed to better DS includes [31], where Wikipedia hyperlinks are used to improve annotation quality. DS can be used to train as many as 5000 different relation extractors [17]. Fine-grained relations can be successfully handled by DS [6]. DS was first applied to German in [5].

All this previous work on DS has in common that knowledge base (structured data) and text (unstructured data) must be in the same language. Our work is the first to show that crosslingual methods can be used to widen the applicability of DS to many additional languages. That allows novel opportunities to improve the quality of DS. Our filtering step is only a simple example of how to use the potential power of crosslinguality for DS.

Significant work on crosslingual resources has been conducted in the Cross-Language Evaluation Forum (CLEF) [30]. Wikipedia is a major resource for this type of crosslingual research. Other relevant crosslingual research includes [1] on generating parallel corpora; [33] on creating a multilingual named entity resource; and [20] on projecting annotations to a target language, a method that, in contrast to crosslingual DS, can only be applied to aligned parallel corpora. Our work for the first time uses crosslingual information for DS.

In [28] a survey is given about crosslingual information extraction (IE) based on SMT. They conclude that there is a significant decrease in accuracy compared to monolingual IE. This is in line with the results of our SMT experiment.

Coverage differences in Wikipedias of different languages using "information arbitrage" is exploited in [2]. The authors train supervised classifiers on parallel structured data to fill in missing information and detect discrepancies between parallel pages. In contrast, we only consider structured data in the pivot language without any need for parallel structured data in the target language. In addition, our approach produces a model that can be applied to any target language text.

7. CONCLUSION

In this paper, we introduced crosslingual DS and evaluated it on four languages. We showed that it has two benefits: (i) higher accuracy of automatic annotation due to CL-filtering and (ii) better coverage because automatic labels can be transferred from the pivot to a target language. Benefit (i) is particularly valuable for complex relations that are modeled in nonrigorous infoboxes. Benefit (ii) applies to both simple and complex relations. We demonstrated that classifiers trained by crosslingual DS are competitive with classifiers trained monolingually on manual labels transferred to the target.

Our long term goal is to use crosslingual DS to automatically generate richly annotated resources for training machine learning algorithms on a large variety of tasks – with relation extraction being only one of these tasks. High accuracy and coverage of crosslingual DS could thus make an important contribution to overcoming the high cost of creating training sets, one of the main impediments to using machine learning more widely in NLP today.

Acknowledgments. We gratefully acknowledge Deutsche Forschungsgemeinschaft (DFG) for funding this research (grant *Sonderforschungsbereich 627*).

8. REFERENCES

[1] S. F. Adafre and M. de Rijke. Finding similar sentences across multiple languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69, 2006.

[2] E. Adar, M. Skinner, and D. S. Weld. Information arbitrage across multi-lingual Wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 94–103, 2009.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference*, pages 11–15, 2007.

[4] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 28–36, 2008.

[5] A. Blessing and H. Schütze. Fine-grained geographical relation extraction from Wikipedia. In *7th international Conference on Language Resources and Evaluation*, pages 2949–2952, 2010.

[6] A. Blessing and H. Schütze. Self-annotation for fine-grained geospatial relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 80–88, 2010.

[7] B. Bohnet. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, 2010.

[8] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1247–1250, 2008.

[9] R. C. Bunescu. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 576–583, 2007.

[10] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Proceedings of the 13th International Conference on World Wide Web*, pages 462–471, 2004.

[11] M. Craven, J. Kumlien, et al. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86, 1999.

[12] A. Culotta, A. McCallum, and J. Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 296–303, 2006.

[13] G. de Melo and G. Weikum. Menta: Inducing multilingual taxonomies from wikipedia. In *Proceedings of the 19th ACM international Conference on Information and knowledge management*, pages 1099–1108, 2010.

[14] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220, 2008.

[15] D. Ferrucci and A. Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10:327–348, 2004.

[16] U. Hermjakob, K. Knight, and H. Daumé III. Name translation in statistical machine translation: Learning when to transliterate. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: HLT*, pages 389–397, 2008.

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

[18] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis. An overview of the TAC2010 knowledge base population track. In *Proceedings of the Third Text Analytics Conference*, 2010.

[19] R. S. Z. Kaljahi. Adapting self-training for semantic role labeling. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 91–96, 2010.

[20] S. Kim, M. Jeong, J. Lee, and G. G. Lee. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 564–571, 2010.

[21] W. Liao and S. Veeramachaneni. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65, 2009.

[22] A. K. McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[23] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1011, 2009.

[24] T.-V. T. Nguyen and A. Moschitti. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 277–282, 2011.

[25] E. Noreen. *Computer-intensive methods for testing hypotheses: An introduction*. Wiley, 1989.

[26] P. V. Ogren, P. G. Wetzler, and S. Bethard. ClearTK: A UIMA toolkit for statistical natural language processing. In *Proceedings of the Workshop on Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*, pages 32–38, 2008.

[27] S. Padó. *User's guide to* `sigf`*: Significance testing by approximate randomisation*, 2006.

[28] K. Parton, K. McKeown, B. Coyne, M. T. Diab, R. Grishman, D. Hakkani-Tür, M. P. Harper, H. Ji, W. Y. Ma, A. Meyers, S. Stolbach, A. Sun, G. Tür, W. Xu, and S. Yaman. Who, What, When, Where, Why? Comparing multiple approaches to the cross-lingual 5W task. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 423–431, 2009.

[29] B. M. Pateman and C. Johnson. Using the Wikipedia link structure to correct the Wikipedia link structure. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 10–18, 2010.

[30] C. Peters, editor. *Cross-Language Information Retrieval and Evaluation, Workshop of Cross-Language Evaluation Forum*, Lecture Notes in Computer Science. Springer, 2001.

[31] S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine learning and Knowledge Discovery in Databases: Part III*, pages 148–163, 2010.

[32] H. Schmid. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the the Association for Computational Linguistics SIGDAT-Workshop*, pages 47–50, 1995.

[33] C. Silberer, W. Wentland, J. Knopp, and M. Hartung. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *6th international Conference on Language Resources and Evaluation*, pages 3230–3237, 2008.

[34] R. Snow, D. Jurafsky, and A. Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, pages 1297–1304, 2004.

[35] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *16th international World Wide Web Conference*, pages 697–706, 2007.

[36] G. Wang, Y. Yu, and H. Zhu. PORE: Positive-only relation extraction from Wikipedia text. In *Proceedings of the 6th International Semantic Web Conference / 2nd Asian Semantic Web Conference*, pages 580–594, 2007.

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

[38] F. Xu, H. Uszkoreit, and H. Li. A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 584–591, 2007.

[39] L. Yao, S. Riedel, and A. McCallum. Collective cross-document relation extraction without labelled data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023, 2010.

[40] T. Zesch, C. Müller, and I. Gurevych. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 60–66, 2008.

[41] G. Zhou, J. Su, J. Zhang, and M. Zhang. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 427–434, 2005.