Pattern Learning for Event Extraction using Monolingual Statistical Machine Translation

Marco Turchi, Vanni Zavarella, Hristo Tanev
European Commission - Joint Research Centre (JRC), IPSC - GlobeSec
Via Fermi 2749, 21020 Ispra (VA) - Italy
marco.turchi@jrc.ec.europa.eu,
{vanni.zavarella, hristo.tanev}@ext.jrc.ec.europa.eu

Abstract

Event extraction systems typically take advantage of language and domain-specific knowledge bases, including patterns that are used to identify specific facts in text; techniques to acquire these patterns can be considered one of the most challenging issues. In this work, we propose a language-independent and weakly-supervised algorithm to automatically discover linear patterns from texts. Our approach is based on a phrase-based statistical machine translation system trained on monolingual data. A bootstrapping version of the algorithm is proposed. Our method was tested on patterns with different domain-specific semantic roles in three languages: English, Spanish and Russian. Performance shows the feasibility of our approach and its capability of working with texts in various languages.

1 Introduction

Multilingual event extraction task consists of retrieving information about particular facts from text documents in different languages and producing event-description templates, which typically contain slots about event participants, location, time and means.

In this work we use an event extraction system which aims at identifying violent events, man made and natural disasters and humanitarian crises, in title and first sentence of news reports. An event is represented as a template, whose main slots correspond to *event-specific semantic roles*, such as: *event-type*, *killed-victims*, *injured-victims*, *perpetrators*, and others. Slot fillers are typically extracted by matching linear patterns in text. For example, *killed <PERSON-GROUP>* represents a sample pattern for the semantic role *DEAD-VICTIM*. It will match text snippets like *killed five people*, where *five people* fills the pattern slot *<PERSON-GROUP>*[1]. In this paper, we are concerned with surface-level, one-slot patterns which accept as slot fillers person names/descriptions such as *two Italian women*.

Building a lexicon of linear patterns is a crucial step in the development and customization of an event extraction system, particularly in news texts which are characterized by an open domain and a large vocabulary. Different approaches have been proposed but most of them require a large manual effort and linguistic expertise. Moreover, due to lexical and syntactic variability and to Zipf's law-based word distribution in language, acquired patterns can only partially cover the range of linguistic constructions. These are serious obstacles faced by every effort to adapt an event extraction system across domains or languages.

In order to address these problems, we put forward a novel language-independent and weakly-supervised algorithm to automatically learn linear event extraction patterns from an unannotated corpus of texts. The method allows knowledge-poor pattern acquisition without any data annotation. It is based on the noisy-channel model developed for Phrase-Based Statistical Machine Translation (PBSMT).

For a particular event-specific semantic role (e.g. *DEAD-VICTIM*) a pattern is proposed as seed. The most frequent person group fillers are selected both automatically from a document collection running an event extraction grammar (Tanev et al., 2009) or manually. Then, a monolingual PBSMT system, separately trained on pairs of comparable sentences from the same language, is used to translate the associations: filler-seed. The new patterns are extracted from the top translations using the mean reciprocal rank (Voorhees,

[1] Notice that *"X pattern"* and *"pattern X"* are two different patterns, with *X* occupying a different position wrt the pattern.

Proceedings of Recent Advances in Natural Language Processing, pages 371–377,
Hissar, Bulgaria, 12-14 September 2011.

2000). This process is bootstrapped passing iteratively the new patterns and the fillers to the algorithm.

Such an approach depends on availability of a corpus of monolingual sentence pairs conveying approximately the same information. The solution we explore is to use pairs composed by the title and the first sentence of a news article. The main idea is that they report about the same content expressed in different ways. A PBSMT system trained on this data produces, as output of the translation process, lexical or morphological variations of the initial seed.

Our algorithm was tested on three languages, namely English, Spanish and Russian, belonging to three different language groups. Manual and application-based evaluations show the adaptability of our approach across languages and domains.

2 Related Work

Systems for automatic event detection and extraction typically use some form of language and domain-specific patterns. Many event extraction systems use syntactic patterns, (Riloff, 1993), or combinations of patterns and statistical classifiers, (Grishman et al., 2005). In the multilingual context, where syntactic parsers are not always available, automatically learned linear patterns are an important resource for event detection and can reach a reasonable level of performance, as shown in (Tanev et al., 2009).

The first pattern learning systems, such as CRYSTAL, (Soderland et al., 1995), and AutoSlog (Riloff, 1993), use manually-annotated corpora. (Riloff, 1996) proposes a weakly supervised method which is an improved version of AutoSlog. This method requires as input a set of text documents, which are manually classified as relevant or irrelevant to a topic. Although this is less demanding than annotating the document content, it is still a time-consuming task. Weakly supervised methods, reported so far, require much less human input than annotating a corpus, but they strongly depend on linguistic knowledge, preventing them from easy adaptation between domains and languages.

Relevant to our work, the multilingual weakly supervised approaches, (Tanev and Wennerberg, 2008) and (Tanev et al., 2009), are based on annotation propagation in semantically consistent document clusters. They share some features with our approach: they use bootstrapping; they only weakly depend on the language; they are domain independent. The disadvantage of these approaches is that clustering is computationally expensive, which prevents this method from scaling to very large corpora.

Another research area, significant to our work is the unsupervised discovery of paraphrases. (Barzilay and Lee, 2003) proposes an approach, which is based on aligned comparable corpora. Unfortunately, such corpora are not easy to be acquired, especially in the multilingual context. In order to go around this obstacle, some approaches use distributional similarity for paraphrase acquisition: For example TEASE, (Szpektor et al., 2004), learns syntactic patterns which paraphrase a seed pattern, but it uses a full syntactic parser, thus making not applicable in a multilingual context. A language independent algorithm to paraphrase English sentences using a Statistical Machine Translation (SMT) system is proposed by (Quirk et al., 2004), where training data are extracted from Web pages and parallel sentences identified using edit distance.

Compared to the aforementioned approaches, our algorithm is more adaptable across languages, since it does not use any language-specific processing. Moreover, our training corpora are easy-to-acquire and more focused on the type of text analysed by the event extraction system, which allowed us to significantly extend training data sets compared to other algorithms based on monolingual machine translation.

3 Monolingual Phrase Based Statistical Machine Translation

Phrase Based Model (Koehn et al., 2003) is an extension of the noisy channel model, introduced by (Brown et al., 1994), using phrases rather than words. The best translation $\hat{e}$ of a source sentence f is obtained by maximizing the probability $p(e|f)$ computed by the product of three components: ϕ, the probability of translating a source phrase f into a target phrase e, d, the distance-based reordering model that drives the system to penalize significant reordering of words during translation and, p_{LM}, the language model probability which assigns a higher probability to fluent/grammatical sentences. Different weight can be associated to each component. For more details see (Koehn et al., 2003). Probabilities are estimated counting the frequency of the phrases in the parallel corpus.

In classical PBSMT, a system is trained using parallel data: each sentence in a source language is associated with correctly translated sentence in a target language. In our approach, we use monolingual comparable data: source and target sentences are respectively the first sentence of the body and the title of a news article in a selected language, for example:

First Sentence: *Twenty-five people were killed in the latest round of Afghan violence this week.*

Title: *25 civilians dead as Taliban intensifies attacks in Afghanistan.*

The main idea is that the two sentences convey the same information in different style, e.g. *Twenty-five people were killed* and *25 civilians dead.* This is grounded on a well-established news writing practice, the so called "inverted pyramid" method, which suggests to re-state the core factual content of a news story at the opening of the article body, (Bell, 1991).

Consequently, a translation in our monolingual PBSMT consists of finding the most probable sentence in the "title" style that contains the same information of the input sentence in the "content" style. In this work, the PBSMT technique allows the extraction of patterns that are indistinctly constituted by either a sequence of words (phrase) or a single word.

4 Pattern Learning Algorithm

The proposed method for pattern acquisition consists of two parts. The first one is the core algorithm with an initial pattern (seed) and a set of fillers, produces a set of reliable new linear patterns. To increase the number of patterns, the core algorithm is then embedded in a bootstrapping schema where it is repeatedly called. In the next Sections, these methods are described in detail.

Core Approach The basic algorithm takes advantage of a monolingual PBSMT system to find lexical and morphological variants of a seed and it is made of three phases: *association*, *translation* and *recombination*. In the first phase, see Fig. 1.a, a set of associations is created pairing the seed, *X killed*, with a set of person/person group fillers, *soldiers*, ... *policemen*, which can be either provided manually or extracted by a person recognition grammar. Each single association is passed to a monolingual translation system, see Fig. 1.b, that produces the top fifty best translations of the association ranked according to $p(e|f)$.

Each seed could be translated by itself, independently from the fillers. However, some initial experiments showed that the filler text snippets help the algorithm to contextualize the translation, e.g. *shot X* with the filler *civilians* or *pictures*. Without any person group context, the extracted variants may end up covering different meanings. Furthermore, filler position crucially defines who or what is doing or undergoing an action in transitive verb group patterns (e.g. *A soldier shot* or *shot a soldier*) so that translating them alone can generate patterns with event roles in inverted position.

In terms of machine translation, the usage of the person group requires the translation of the full association, person group plus seed, rather than using the translation model as a look-up table for the seed only. This means that the SMT may also produce a variation of the person group adding extra noise to the output. To reduce the impact of the presence of the person group, each association is passed to the SMT system with an option that forces PBSMT not to modify the filler in the output, but to use it to contextualize the translation, e.g. *soldiers* and *policemen* are present in their original form in Fig. 1.b.

For a single seed, sets of translations are generated according to the number of associations, and the same new pattern can be ranked in a different position inside different sets, e.g. *are killed* as shown in Fig. 1.b. The last step consists of extracting all the new patterns from the sets of translations and re-ranking them in a reliability order, see Fig. 1.c. To make the new patterns comparable across sets, in each translation the person group is substituted by a X. The recombination of the patterns in a unique list can be done merging and re-ranking all of them using a mathematical operator based on $p(e|f)$, e.g. average, but $p(e|f)$ is a local property of each set of translations because includes the contribution of the filler.

The main idea that we propose consists of using a global metric that takes advantage of the local rank inside of each set. For this purpose we use the Mean Reciprocal Rank (MRR), (Voorhees, 2000), a metrics used in information retrieval to evaluate any process that produces a list of possible responses to a query. The mean reciprocal rank for a sample of queries Q is reported in Eq. 1 where

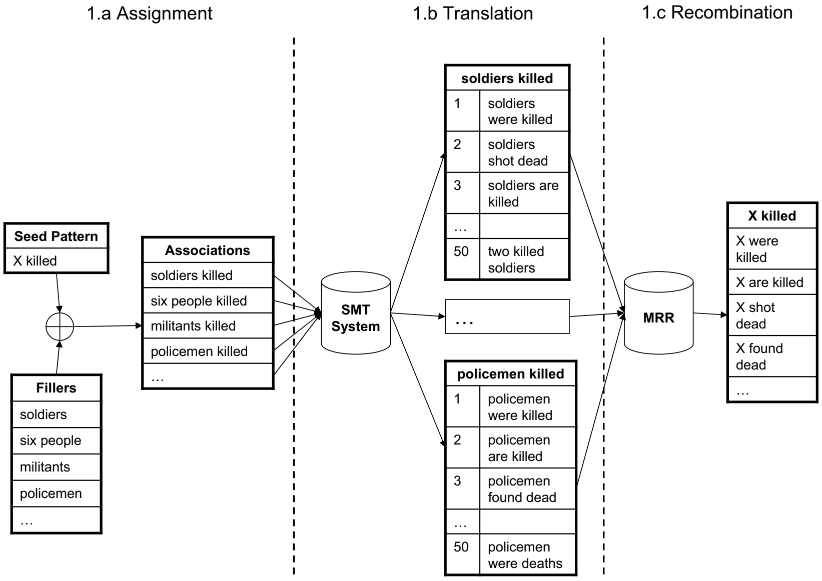

Figure 1: Extraction of new patterns using the seed "*X killed*".

$rank$ is the rank of the first correct answer.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i} \quad (1) \qquad MRR(np) = \frac{1}{|A|}\sum_{i=1}^{A}\frac{1}{rank_i(np)} \quad (2)$$

We adapt the MRR in the following way: the number of queries is the number of associations, and the answers to a query are the fifty translations of a certain association. The MRR of a new pattern, np, is shown in Eq. 2, where A is the number of associations for a seed and $rank_i(np)$ is the position of np in the set of translations of the association i.

High rank of a new pattern in a set of translations guarantees its correctness while MRR based on the ranked translations guarantees that those new patterns that are on top positions in various sets received a high rank in the final list. Top patterns are selected from the final list by picking up those that have MRR value bigger than 10% of the MRR value of the best pattern.

Bootstrapping. The core algorithm is embedded in a bootstrapping framework. Starting from the original seed, called "root seed", each new pattern produced by the core algorithm can be considered an input seed for another instance of the core algorithm. This procedure can be iterated over all the new patterns.

This approach increases the number of retrieved patterns, but can create unwanted noise. At each bootstrapping step, the produced patterns can be semantically divergent from the "root seed" because a seed can be semantically ambiguous or polysemous and one of the fillers can be too generic to pick up a unique sense. We tackle this problem by introducing a stop criterion in the bootstrapping framework, whose goal is to select only those new patterns that are semantically similar to the "root seed". The selected patterns are only propagated to the next iteration and the seed that produced them is added to the final results.

The concept of semantic similarity between a surface pattern and the seed is modelled by simply using set intersection.We assume that the new patterns produced by the expansion of the "root seed" are the most semantically similar to it. This is confirmed by the value of the macro-precision of the produced patterns in English which is equal to 82%. According to this, we consider the expansion of the "root seed" as a gold standard, GS, for the bootstrapping approach.

At a certain bootstrapping step t, a new pattern, np^t, is passed to the core algorithm, CA, producing a set of new patterns, $S(np^t)$. np^t is semantically correlated to the "root seed" if and only if the intersection between GS and $S(np^t)$ is not empty. It means that $S(np^t)$ should have at least one new pattern in common with the gold standard for being semantically correlated to the "root seed". If the condition is true, np^t is added to the final results and the new patterns, that are not in common with the gold standard, are propagated to next bootstrapping iterations. Otherwise the bootstrapping is stopped, np_t is not considered a reliable pattern and not included in the final results. The stop criterion forces each new pattern at iteration t to be validated using information produced at iteration $t + 1$ before being added to final results.

The stop criterion is not highly restrictive, it reduces the number of computations and guarantees a semantic similarity between the "root seed" and new patterns. The final output of the bootstrapping process is the union without duplicates of all the new patterns that are evaluated as correct by the stop criterion.

5 Experimental Setup

In this work, we use Moses, (Koehn et al., 2007), a complete phrase-based translation toolkit for research purposes.

Training data are extracted from a title and first sentence of news articles gathered during a one year time span from 01/07/2008 to 01/07/2009. We perform experiments in three languages: English, Russian and Spanish. For each language, we respectively use ∼2,87M, ∼2,19M and ∼1,48M sentence pairs. Nine event predicates are chosen, which are important for analysis of political, crisis and violence-related news (for example *DEAD-VICTIM*). For each of them a highly frequent and unambiguous linguistic realization is selected as a single-slot seed pattern, for each of the three languages: *X sentenced* (1), *criticized X* (2), *X visited* (3), *X were killed* (4), *X met with* (5), *X were evacuated* (6), *X were wounded* (7), *supported X* (8) and *X launched an attack* (9) [2]. In each language, seed patterns are integrated with person/person group recognition rules, as proposed in (Tanev et al., 2009), and run on a news corpus to extract a set of person/person group fillers: the 20 most frequent are then paired with the seed pattern and fed to the PBSMT system[3].

6 Evaluation and Results

We evaluate by running only four iterations of bootstrapping, where the fourth is used to validate the new patterns extracted at iteration 3. An average of about 55, 74 and 39 new patterns over all the predicates are acquired for English, Russian and Spanish, respectively. There were rates of 3.6%, 0.3% and 4.8% ungrammatical patterns. For a seed that was not in the test set, *X was kidnapped*, we experimented running more iterations of bootstrapping, finding that at each iteration the number of correct patterns grew about 1.5 times on average, at the cost of a small decrease of precision (about 20%). The number of new patterns is relatively small, because we wanted to test the generative power of the algorithm when fed with a minimal input of only one seed pattern.

We performed a direct evaluation of the output pattern Accuracy and then we evaluated indirectly the Precision and Recall via running an extraction system. Ungrammatical patterns are considered inapplicable and discarded from accuracy evaluation while we keep them for evaluating extraction performance.

Pattern Accuracy. Pattern Accuracy evaluation was performed by asking a language expert to rate each pattern as either "correct" (semantically sound and non-ambiguous), "correct-in-context" (partially ambiguous but semantically sound in some linguistic context) or "incorrect". A "lenient" Accuracy score was computed as the ratio of both the "correct" and "correct-in-context" patterns over the total, while "strict" accuracy only includes "correct" patterns.

Id	**English**		**Russian**		**Spanish**	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
1	0.42	0.66	0.68	0.70	0.32	0.36
2	0.43	0.61	0.00	0.00	0.19	0.29
3	0.51	0.78	0.23	0.33	0.29	0.40
4	**0.64**	0.81	0.52	0.64	**0.83**	**1.00**
5	0.57	0.80	**0.76**	**0.87**	0.77	0.77
6	0.59	**0.83**	0.50	0.64	0.26	0.30
7	0.35	0.47	0.30	0.34	0.57	0.57
8	0.31	0.37	0.62	0.85	0.31	0.38
9	0.61	0.69	0.48	0.60	0.00	0.00
	0.49	**0.67**	**0.46**	**0.55**	**0.39**	**0.45**

Table 1: Manual evaluation of pattern accuracy. Highest values are highlighted.

Average Kappa score between two annotators over the 9 pattern sets for English was 0.58, which is in the higher range "moderate agreement" class according to (Fleiss, 1981). However, the Kendall tau-b rank correlation coefficient, (Lapata, 2006), turns out to be a more suitable evaluation metrics as it better accounts for the natural ordering of the rank classes. We measured a 0.79 score ($p < 10^{-3}$), consequently we assumed the anno-

[2] In the next Sections, we refer to each pattern using the number close to it.

[3] Notice that fillers could have been manually produced as well, so that the overall algorithm is not really dependent on the person recognition grammar.

tation task is grounded and performed it with one single annotator for Spanish and Russian.

Pattern Accuracy scores for each predicate are shown in Table 1 together with macro-averages. Among correct patterns in all the seeds, morphological variants can be observed (including mood, tense, number) as well as lexical shifts and a few verb form alternations (e.g. active-passive). A common source of noise is the assignment of the filler position to a wrong verb argument (e.g. *X were killed → X kills*; *supported X → X favour*). This is due to the reordering model in the PBSMT system that considers the incorrect position of the filler as probable as the correct one, so forcing the translation system to output the wrong pattern.

Overall, pattern Accuracy figures closely correlates with the size of the training corpora for the PBSMT systems in the three languages.

Extraction systems based on the same schema (initial seed plus bootstrapping approach in a unsupervised manner) have accuracy on new patterns from 40% to 50% (e.g. the Web-based system by (Szpektor et al., 2004)), consequently we consider the performance of our method for pattern learning really encouraging.

Event Extraction Precision and Recall. In order to measure Recall and Precision of the new pattern sets, we compared performance of a baseline extraction system (**BL**), containing person entity grammar and the single seed extraction pattern, against a target system (**TG**), that adds the set of the discovered patterns to the seed, and then against a clean target system (**CT**), that adds only those discovered patterns that are human-evaluated as "correct" and "correct-in-context".

Recall was measured in the following way: for each event predicate, a set of 20 news article sentences reporting about that event type were manually collected then the person/person group entity expressions were replaced in text with a constant expression detectable by the person recognition grammar, so as to make the results unaffected by the performance of the grammar itself. Then the number of successful detections of that filler was checked.

As for the Precision, the baseline and target systems were both run on a corpus of titles and first sentences of news articles collected during 10 days, resulting in about 5.79M, 3.29M and 700k words for English, Russian and Spanish respectively. From all the system outputs, a set of 20 were randomly collected, discarding duplicates, and the correctness of extracted fillers were manually evaluated. Answers were rated as correct when at least one of the fillers extracted was at least partially overlapping with the full person entity expression actually in text [4].

Table 2 shows Precision and Recall scores of the discovered patterns in an extraction task[5]. The Recall of the TG system is raising constantly from the baseline values across all the predicates and for each language. Recall can be improved raising the number of correct patterns added to the system. This, as mentioned in Section 6, can be done by increasing the number of bootstrapping steps. Precision of the TG system is also constantly dropping. However, this decrease can be significantly limited via human pattern selection, as can be seen from the performance of the CT. system Overall, the automatic approach proposed here, coupled with a lightweight human post-processing step, generates a good quality pattern lexicon for information extraction.

For the TG system, performance seems to be largely variable across predicate types, and this partially correlates with the pattern accuracy figures too. However, performances seem to be independent from domain variation, with the best results spreading over the violent, political or judicial event domains. This suggests that domain adaptation of an event extraction system can be easily achieved in our method by providing a suitable amount of training data in the corresponding subject domain, so as to reduce the ambiguity of the language.

7 Conclusions

We proposed a language-independent and weakly-supervised bootstrapping algorithm to learn linear patterns from text, based on a phrase-based statistical machine translation system trained on monolingual data.

Among the different methods that have been proposed for extracting linear patterns from text, our approach is completely language independent, and it relies on freely available data such as news articles. Training data for the SMT system do not require any heavy pre-processing and such sen-

[4] E.g. *soldiers* is taken as a correct system answer for the *injured-victim* role in a sentence like "*3 German soldiers were wounded*"

[5] F-measure scores could not be computed on such Precision and Recall figures coming from different test sets

Id	English						Russian						Spanish					
	P			R			P			R			P			R		
	BL	TG	CT	BL	TG	CT	BL	TG	CT	BL	TG	CT	BL	TG	CT	BL	TG	CT
1	0.90	0.40	**0.85**	0.10	0.30	0.30	1.00	0.40	0.80	0.10	0.50	0.30	na	0.00	0.25	0.00	0.36	**0.35**
2	1.00	0.50	0.40	0.25	0.45	0.45	0.00	0.50	na	0.00	0.00	0.00	na	0.00	0.25	0.05	0.20	0.05
3	0.90	0.30	0.60	0.10	0.40	0.40	0.95	0.30	0.85	0.10	**0.70**	**0.65**	1.00	0.60	0.70	0.20	0.30	0.35
4	1.00	0.60	0.65	0.00	0.30	0.30	1.00	**0.60**	0.90	0.25	0.80	0.80	1.00	**1.00**	**1.00**	0.10	0.15	0.15
5	0.95	**0.60**	0.50	0.00	0.40	0.30	1.00	**0.60**	**0.95**	0.10	0.45	0.40	1.00	**1.00**	**1.00**	0.00	**0.50**	0.05
6	0.93	0.25	0.55	0.05	0.30	0.30	0.95	0.25	0.80	0.00	0.45	0.45	1.00	0.10	**1.00**	0.00	0.05	0.05
7	0.90	0.05	0.45	0.00	**0.70**	0.65	1.00	0.05	0.80	0.00	0.30	0.05	1.00	0.30	**1.00**	0.05	0.05	0.05
8	0.50	0.15	0.30	0.00	0.80	**0.80**	0.64	0.15	0.45	0.10	0.55	0.55	na	0.35	0.30	0.00	0.15	0.25
9	1.00	0.25	0.35	0.00	0.35	0.35	1.00	0.25	0.70	0.00	0.50	0.35	na	na	na	0.00	0.00	0.00
	0.90	**0.34**	**0.52**	**0.06**	**0.43**	**0.38**	**0.84**	**0.34**	**0.78**	**0.07**	**0.47**	**0.39**	**1.00**	**0.42**	**0.69**	**0.04**	**0.20**	**0.14**

Table 2: Pattern performance in an extraction task. "na" values for Precision mean that there were no extracted fillers for that test set. For each language, the biggest improvement (or smallest decrease) over the pattern types compared to the baseline is underlined.

tence pair collections can be easily built for any language and target domain from the news.

The new extracted patterns, in the "title" style, contain exactly the kind of variation in linguistic constructions that the event extraction system has to deal with during the detection process on title and first sentence of a news article. Performance analysis confirms this assumption and shows the feasibility of the approach both across languages and domains.

From an evaluation of the output patterns we noticed a degradation of the Accuracy after the first iteration of the algorithm. It is our intention to investigate the role of the bootstrapping criterion and model the similarity condition with some robust measure of distributional similarity between pattern sets.

Acknowledgments

Special thanks go to Francesca Torti from University of Parma for her essential statistical support.

References

Barzilay R., and Lee L. (2003) Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. *Proceedings of HLT-NAACL*, 16–23. Edmonton, Canada.

Bell A. (1991) *The Language of News Media*. Blackwell Publishers, Oxford.

Brown P.F., Della Pietra S., Della Pietra V.J., and Mercer R.L. (1994) The Mathematic of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2) 263–311.

Fleiss J.L. (1981) *Statistical methods for rates and proportions*. John Wiley, New York, USA.

Grishman R., Westbrook D., and Meyers A. (2005) NYU's English ACE 2005 system description. *Proceedings of the ACE, Evaluation Workshop*.

Koehn P., Och F.J., and Marcu D. (2003) Statistical phrase-based translation. *Proceedings of NAACL*, 48–54, Morristown, USA.

Koehn P., Hoang H., Birch A., Callison-Burch C., et al. (2007) Moses: Open source toolkit for statistical machine translation. *Proceedings of ACL*, 45(2).

Lapata, M. (2006) Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4) 471–484.

Quirk C., Brockett C., and Dolan W. (2004) Monolingual machine translation for paraphrase generation. *Proceedings of EMNLP*, 149. Barcelona, Spain.

Riloff E. (1993) Automatically Constructing a Dictionary for Information Extraction Tasks. *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 811–816. Seattle, USA.

Riloff E. (1996) Automatically Generating Extraction Patterns from Untagged Text. *Proceedings of AAAI*, 1044–1049. Portland, USA.

Soderland S., Fisher D., Aseltine J., and Lehnert W. (1995) CRYSTAL: Inducing a conceptual dictionary. *Proceedings of IJCAI*, 1314–1319. Canada.

Szpektor I., Tanev H., Dagan I., and Coppola B. (2004) Scaling Web-based Acquisition of Entailment Relations. *Proceedings of EMNLP*, 41–48. Spain.

Tanev H., and Wennerberg P. (2008) Learning to Populate an Ontology of Politically Motivated Violent Events. *Mining Massive Data Sets for Security*, IOS Press, 311–322.

Tanev H., Zavarella V., Linge J., Kabadjov M., et al. (2009) Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. *Linguamatica*, 2(1) 55–66.

Voorhees E.M. (2000) The TREC-8 question answering track report. *NIST Special Publication*, 77–82.