IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 27-43
RECEIVED 27/11/09 ACCEPTED 16/01/10 FINAL 10/03/10

Learning Event Semantics From Online News

HRISTO TANEV, MIJAIL KABADJOV, AND MONICA GEMO

Joint Research Centre, Italy

ABSTRACT

In this paper we present a multilingual algorithm for automatic extension of an event extraction grammar by unsupervised learning of semantic clusters of terms. In particular, we tested our algorithm to learn terms which are relevant for detection of displacement and evacuation events. Such events constitute an important part in the process of development of humanitarian crises, conflicts and natural and man made disasters. Apart from the grammar extension we consider our learning algorithm and the obtained semantic classes as a first step towards the semi-automatic building of a domain-specific ontology of disaster events. We carried out experiments both for English and Spanish languages and obtained promising results.

1 INTRODUCTION

Automatic event extraction is a relatively new sub-branch of information extraction, whose ultimate goal is the automatic extraction of structured information about events described in text sources, such as news. We look at the events as complex processes including interactions among several entities. Each of these participating entities has an event-specific semantic role, which defines the way in which the entity participates in the event and interacts with the other entities. The event-specific semantic roles are related to the nature of the entities to which they are assigned, however this relation is not straightforward. For example, in the context of evacuation events, an entity which belongs to the category *buildings*

can be assigned the event-specific role *evacuated-place*, *target-place-of-evacuation* (a place where people are evacuated) or it can be related only loosely to the dynamics of the event.

In the context of our preceding and current work, this relation is modeled through event extraction grammars which connect linguistic expressions, such as "were evacuated" with semantic classes, such as *person-group*, *facility*, etc. and event-specific semantic roles, such as *evacuated-people*, *evacuated-place*, etc. For example, the following sample rule detects evacuation events and extracts descriptions of evacuated people and evacuated places:

person-group :evacuated-people *were evacuated from* **NP(**head-noun: *place*): evacuated-place

This rule will match a text like: "Five women were evacuated from a hotel." and will extract "five women" as *evacuated-people* and "a hotel" as *evacuated-place*.

In order to automatize partially the process of creation of such rules, we propose a semi-automatic approach for extending event extraction grammars. The core of our approach is an unsupervised algorithm for learning of semantically consistent term clusters. In particular, we tested our algorithm to acquire terms, which are relevant for detection of displacement and evacuation events. Such events constitute an important part in the process of evolution over time of humanitarian crises, conflicts and natural and man made disasters. Apart from the grammar extension, we consider our learning algorithm and the obtained semantic clusters as a first step towards the semi-automatic building of domain-specific ontology of disaster events.

The starting point for us is an existing event extraction grammar for detection of evacuations and displacements from online news reports. The grammar is an integral part of NEXUS [1], an automatic system for event extraction from online news, which is profiled in the domain of security and crises-management.NEXUS makes use of over 90 event-specific patterns and a noun-phrase recognition grammar to detect boundaries of phrases which refer to groups of people.Using these two resources the system can identify text fragments such as "about 200000 people have abandoned their homes", where the phrase "about 200000 people" will be labeled with the event-specific semantic category *displaced people*.

Similarly, NEXUS can identify phrases about evacuations, such as "five women were evacuated", where "five women" will be labeled as *evacuated people*.

However, further, crucial information about these event scenarios is often encoded by the subcategorization frames in which the main verb phrases of the patterns may occur or by some verb adjuncts. In both cases, they consist of highly productive prepositional phrases (with selectional restrictions), where the noun-phrase head typically belongs to a specific semantic category. For example, in the text fragment "more than 1000 people were evacuated after a chemical leak" the prepositional phrase contains the crucial information about the event which caused the evacuation. In a similar way, the phrase "20000 people displaced to Beddawi camp" reports both the number of displaced people as well as the place where they were moved.

We developed an algorithm which expands automatically the event extraction grammar by learning a subset of the scenario-related subcategorization frames of verb phrases from unannotated news corpus.

The main part of our learning algorithm is an unsupervised term extraction and clustering approach which is a new way of combining several state-of-the-art term acquisition and classification techniques.

Clearly, there is far more structure within the subcategorization frames of the domain-specific verbs than standard surface level patterns of NEXUS can detect. Consequently, more work will be necessary to obtain a better picture about the different syntactic positions in which the semantic clusters can be introduced with respect to the main verbs. At this stage, we regard our experiments just as a first step towards automatic or semi-automatic learning of syntactico-semantic rules.

The rest of the paper proceeds as follows: Section 2 makes a review of the related work. Section 3 introduces the event extraction grammar, currently exploited by NEXUS. Section 4 explains our approach for learning of semantic classes and extending the event extraction grammar. Section 5 describes our experiments and the evaluation we did. Finally, section 6 presents our conclusions and discusses future research directions.

2 RELATED WORK

Relevant to our work are approaches for learning of verb subcategorization frames. In particular the work of [2] share similarities with our method, as far as it is based on automatic term clustering to acquire semantic clusters in unsupervised manner. However, they rely on manual

attachment of the obtained semantic clusters to the prepositions. Moreover, their semantic clusters are of limited size and contain only nouns which appear with specific predicates. Apart from application on very specific domains, such an approach will require a large training corpus to compensate for potential data-sparseness problems.

Another group of approaches in this field were introduced by the work of [3]: They use thesauri or taxonomies, such as WordNet to find the right level of semantic generalization in the subcategorization frames. The problem is that these methods are hardly applicable for languages other than English, due to the extensive use of semantic resources

Clustering and classification of words based on the distributional similarity of their contexts is not new: [4] proposed this approach for automatically clustering of nouns. Later, [5] used different syntactic features to cluster semantically similar words. Recently, the interest to distributional-similarity approaches was revived in the context of Ontology Learning and Population - [6] introduced unsupervised approach for ontology population, based on context distributional similarity between named entities, such as “Trento” and semantic categories, such as “city”; based on this work, [7] introduced some limited-scale supervision in the form of semi-automatically acquired seed sets of named entities thus improving the performance. The approach, presented by [8] uses contextual based similarity to cluster words into concept clusters; a particular feature of this work is that it explores deeper the usage of concept attributes such as contextual features. The problem with these approaches is that they begin with a predefined set of terms, taken from ontologies or other sources, which are next clustered or classified. It is not clear what will be the performance when combined with term extraction from free texts.

Another type of approaches for semantic classification follow the pioneering work of Marti Hearst [9]. It puts forward a small set of hypernym-hyponym extraction patterns, which relate a concept word, such as “city” with its possible hypernyms, e.g. “place”. A similar pattern-based approach was used by [8] to extract concept attributes. However, such pattern-based approaches are strongly affected by the data sparseness problem (see [7]), some authors promote the use of the Web [10] via a search engine, which however brings under consideration problems such as efficiency, maximal number of allowed queries, access policies of the search engines, etc.

3 EVENT EXTRACTION GRAMMAR

The grammar currently used by NEXUS is a finite state cascade grammar. The first grammar level recognizes references to groups of people, such as "100 women", "fifty Chinese workers", etc. The first level works on the top of a tokenizer and a dictionary with person referring nouns, such as "women", "workers", etc., nations, such as "Chinese", "Russian", etc. As an example, consider the following grammar rule which can parse phrases like "5 Canadian soldiers":

[person-group] → **(digit-number** | **word-number+) nation? person-noun-plural**

The second grammar cascade combines the recognized person phrases from the first level with the linear patterns, listed in a dictionary. As an example consider the following patterns for recognition of displacement events:

[person-group] *were forced out of their homes*

[person-group] *were displaced*

[person-group] *were uprooted*

These patterns and others of their type are encoded at the second grammar level through one rule:

[person-group] right-context-displacement-pattern

In this rule *right-context-displacement-pattern* refers to a class of string patterns, listed in the pattern dictionary, such as "were displaced", "were uprooted", etc. These strings, when appearing on the right from a description of a person group, designate a description of displacement event, in which the person group phrase refers to the displaced people in this event

4 EXTENDING THE GRAMMAR

The goal of the grammar expanding algorithm is the learning of syntactic adjuncts which are introduced in the description of the events usually

through prepositional phrases. More concretely, we would like to recognize phrases like "many people were evacuated to temporary shelters". In order to do this, our system has to recognize patterns like

[person-group]*were displaced to* **NP**(*facility*)

where *NP(facility)* refers to a noun phrase whose head noun belongs to the category *facility*, which should be described through a list of nouns. The grammar should also assign the event-specific semantic label *place-of-displacement* to this noun phrase. In the context of our experiments, we learn grammar extensions in the form of triples *(preposition, semantic cluster, event-specific role)*. For example, *(to; F; place-of-displacement)*, where F is a cluster of words, which can be considered as belonging to the category $facility$ in our event specific context.

We do not specify which triple to which pattern can be attached. This was not done, since many patterns are based on the same verbs or at least on verbs which share the same or similar sub-categorization frames. Therefore, the sample triple , *(to; F; place-of-displacement)* will be encoded in the extended grammar as

[person-group] right-context-displacement-pattern *to* (**NP**(F)):**place-of-displacement**

left-context-displacement-pattern [person-group] *to* (**NP**(F)):**place-of-displacement**

where F refers to a cluster, represented via dictionary which contains words which are likely to be $facilities$, e.g. "school", "hospital", "refugee camp", etc. Such rules can recognize text fragments, such as "1000 people were displaced to government shelters", provided that "shelters" is a member of the cluster F. Moreover, "government shelters" will be tagged with the event-specific semantic labels *place-of-displacement*.

4.1 *Algorithm overview*

In order to learn such grammar extensions, we propose the following multilingual machine learning algorithm, on which we elaborate in the following subsections:

1. Create a superficial seed terminology extraction grammar which recognizes preposition phrases which appear after displacement/evacuation patterns. We obtained this grammar via extending the multilingual

event extraction grammar of NEXUS. Note, that this is NOT the final extended grammar which was discussed in the beginning of this section, although its structure is very similar. It is rather a grammar for extraction of seed terminology.
2. We run the term extraction grammar on a news corpus and we extract all the pairs of a preposition and a head noun which appear after the event extraction templates. These pairs are grouped by preposition. In such a way, we obtain for each preposition a list of nouns which appear after it.
3. For each preposition, we cluster the corresponding nouns, using distributional similarity of their contexts.
4. We extend the clusters, using a multilingual term extraction based on context distribution similarity.
5. Clusters are cleaned using Hearst hypernym-hyponym templates applied on the Web.
6. Manually, we link each learned pair of a preposition and a semantic cluster to an event-specific semantic role, such as *cause-of-displacement*.
7. Extend the event extraction grammar by adding the learned adjuncts

4.2 *Seed terminology extraction grammar*

We construct the term extraction grammar by extending the second level event extraction grammar, described in the previous section. We created simple noun phrase recognition rules which utilize the output of a morphological processor. These rules constitute an intermediate grammar level between the first and the second one. The output of this level is the structure $NP(head : N)$, which denotes a noun phrase with head N. The second level rules for displacement and evacuation are modified by adding an adjunct introduced by a preposition. For example,

[person-group] right-context-displacement-pattern

will become

[person-group] right-context-displacement-pattern Prep NP

where $Prep$ can match any preposition and *NP* matches any noun phrase. Similarly, the preposition and the NP are attached to left context rules and the same we do for evacuation patterns. This grammar is used during the learning phase to extract a list of terms which are next used to form seed semantic classes.

4.3 *Learning semantic classes*

We run the term extraction grammar on a news corpus and we extract all the pairs of a preposition and a head noun which matches the construction *Prep NP(head-noun:n)*. If the NP has a main noun modified by another noun, e.g. "rain fall", then the whole bi-gram is taken as a head noun. As an example, from the text "five people were evacuated from a burning hotel", the term extraction grammar will extract the pair ("from", "hotel")

All the extracted pairs are grouped together with respect to the preposition. For each preposition, we keep only these nouns which appear with it at least a certain number of times. In such a way we obtain a list of prepositions, and for each preposition we have a list of associated nouns (or noun bi-grams, as explained before). For example, let's assume that for the preposition "after" we obtain the list: "day", "fire", "forest fire", "dam break", "flood", "rainstorm". Then, the following cluster learning algorithm is applied:

1. For each word in a cluster we obtain a list of contextual features. They are uni-grams, bi-grams and tri-grams which co-occur with the word in a news corpus. Weighting is carried out using an algorithm similar to the one described in [7], however we use superficial features, similar to the ones used by [11]. The feature weighting is described in more details in the next subsection.
2. The nouns corresponding to one preposition are clustered based on their contextual features extracted in the previous step. This step is necessary, since the same preposition can be followed by nouns from different semantic classes, which introduce different event-specific semantic roles. For example, the nouns occurring after the preposition "after" are clustered in three seed clusters:
   - day
   - fire, forest fire
   - dam break, flood, rainstorm
3. We ignore seed clusters with less than 3 elements as unreliable, therefore only the third cluster will remain in the previous example. Since clusters are formed based on contextual features, then words in a cluster will tend to appear in similar contexts. According to Harris' distributional hypothesis words which appear in similar contexts have similar semantics.
4. We use each seed cluster as a seed set to learn new terms which have similar contextual features and therefore are semantically similar. We used our in-house term extraction system, opulis, to perform this task. The system is based on a weakly supervised ontology

population approach introduced in [7] and modified for the use with superficial features. From an initial seed set of terms, Ontopopulis learns a list of terms with similar contextual distribution. The list is ordered by similarity with the seed set. We use the first 300 most similar elements from it to form our extended cluster. As an example, consider the highest scored members of the extended cluster obtained from the seed cluster "dam break", "flood", "rainstorm"; the top-scored members of the extended cluster, obtained from it, are: "flood", "quake", "floods", "fire", "tsunami", "disaster", "flooding", "earthquake", "storm", "cyclone", "hurricane", etc.

5. The extended cluster generated by the top 300 elements returned by Ontopopulis have significant amount of noise due to the big number of accepted terms. On the other hand, we found that some correct terms can have low similarity score due to data sparseness, semantic ambiguity, etc. Therefore, reducing the number of the accepted terms would result in low coverage. In order to improve the semantic consistency of our clusters without discarding many appropriate terms, we propose a semantic validation approach, based on superficial hypernym-hyponym patterns, similar to the ones introduced by Marti Hearst in [9]; we used the Web as a corpus. The approach has three main steps:
   - First, for the seed cluster, for example ("dam break", "flood", "rainstorm"), it forms the plural forms of the words: "dam breaks", "floods", "rainstorms" and queries the Web, using Yahoo API, with the pattern "such * as W", where W is substituted with the plural form of each word in the seed cluster, e.g. "such '* as dam breaks". For Spanish we used the pattern "W y otros * ".The assumption is that what appears at the position of the asterisk will be mostly a word X, such that there is an *is-a* relation between the word W and X. That is, X can be considered a hypernym of the word, at least in certain contexts.
   - Next, we learn one hypernym word H which co-occurs with most of the words from the seed cluster. (We use a simple co-occurrence measure based on frequencies). For the example seed cluster we obtain the hypernym word: "disasters"
   - For each word W from the extended cluster the algorithm forms its plural form WP and queries Yahoo API with the check pattern "H such as WP' (for Spanish it becomes "WP y otros H")', e.g. "disasters such as hurricanes". If seven or more pages are found on the Web which contain the pattern, then the word W

is accepted, otherwise it is filtered out from the extended cluster. In such a way we leave in the clusters mostly words which are likely to have an *is a* relation with one and the same concept. This improves semantic consistency of the final cluster. Note, that for English the check pattern is a slightly modified version of the Hearst pattern used to learn a hypernym in the previous step; the motivation for using two patterns is empirical - the first one is more precise and therefore better for learning of hypernyms, however we found it to be too restrictive as a check pattern.

At the end, we link each semantic cluster with the prepositions from which its seed cluster co-occurs. Therefore, at the end of this learning phase we have a list of word clusters $C_1, C_2, ...C_n$, which are mostly semantically consistent and a list of pairs $(Prep, C_i)$, where $Prep$ denotes a preposition and C_i denotes a cluster.

CONTEXTUAL FEATURES The basis of the semantic cluster learning are the contextual features. In our work a contextual feature of a word w is defined to be any lowercase word, bi-gram or a tri-gram which co-occurs in a corpus immediately on the left or on the right from w, it is not a stop-word, and co-occurs at least certain number of times. The co-occurrence feature specifies also the position of the n-gram (left or right) with respect to the words. For example, the word "hurricane" has a feature "X destroyed", where X shows the position of the word (in this case "hurricane") with respect to the feature. Every contextual feature is weighted, based on its co-occurrence with the word. Co-occurrence is measured using the Pointwise Mutual Information. These contextual features were used both for intital word clustering for obtaining the seed clusters, as well as for their expansion with Ontopopulis. When measuring the contextual similarity of two words, the dot product of their feature vectors is calculated.

4.4 *Extending the event-extraction grammar*

As it was pointed out before, at the end of the previous step we obtain a list of semantic clusters $C_1, C_2, ..., C_n$ and a list of pairs $(Prep, C_i)$, where $Prep$ denotes a preposition and C_i denotes a cluster. We manually link each pair to an event-specific semantic role, such as *cause-for-displacement* (e.g. "forest fire"), *target-place-of-displacement, means-of-evacuation, psychological-state-of-evacuated* (e.g., "left the building in

panic"), *evacuated-place* (e.g. "were evacuated from a skyscraper"), etc. In such away, we transform each pair *(Prep, Ci)* into a triple *(Prep, Ci, event-role)*, where *event-role* is a manually-assigned event-specific semantic role, such as *cause-for-displacement*. Each triple $(Prep, Ci, event-role)$ is used to transform each domain specific rule, such as:

[person-group] right-context-displacement-pattern

into

[person-group] right-context-displacement-pattern Token? Token? Token? Prep NP(head-noun:Ci): event-role

The term $NP(head-noun : C_i)$ will match each noun phrase, whose head noun belongs to the cluster Ci. (We used the noun-phrase extraction grammar layer, described in the second subsection.) In order to augment the coverage of the extended rules, we allow for several optional tokens to appear between the original pattern and the prepositional phrase. The *event-role* shows what event specific role will be assigned to the noun phrase, which matches $NP(head-noun : C_i)$. After several experiments with this grammar we reached the conclusion that some semantic clusters nearly always introduce the same event specific role, when appearing closely to the pattern, and this does not depend on the preposition. For example, the cluster with disasters always shows the reason of displacement. In such cases we omit from the rules the specification of the preposition and allow for more optional tokens, which can appear between the pattern and the noun phrase matched by $NP(head-noun : C_i)$. In such a way we increased the generality of our rules and obtained higher coverage for the extended grammar.

5 EXPERIMENTS AND EVALUATION

We applied our algorithm on English language online news, we obtained several semantic clusters, which we used to extend our event extraction grammar and extract three new types of event-specific roles, namely *cause for displacement/evacuation*, *evacuated place* and *target place of the evacuation*. We carried out also experiments with Spanish-language online news. Since we did not have enough time, we did not run the whole learning algorithm for the Spanish. We learned two semantic clusters for

this language and added one new semantic role to the Spanish event extraction grammar, namely *cause for displacement/evacuation*.

5.1 *Experiments for English*

We extended the English language grammar to obtain a term extraction grammar, as explained in section 4.2. The grammar uses 103 patterns for displacement and evacuation events.

Then, we run the algorithm for learning of the semantic classes, described in section 4.3 : We run the event extraction grammar on a $6GB$ corpus of news articles excerpts and extracted 11 prepositions which tend to appear after patterns for evacuation and displacement. For each preposition the event extraction grammar extracted also a list of nouns which tend to appear frequently after it. For our experiments we chose 5 of them for which there were sufficient number of nouns. These were the prepositions: “after”, “to”, “into”, “from” and “in”. For each of them we took the list of nouns and performed agglomerative clustering, based on contextual features, which were extracted from a news corpus. We chose in random a couple of clusters from each preposition; we expanded and cleaned them using the learning algorithm described in section 4.3. In such a way, we obtained 8 semantic clusters. We manually labeled with events-specific semantic roles all, but one of the combinations of preposition-cluster pairs. We used three types of event-specific semantic roles: *cause for displacement/evaluation*, *source (evacuated place)* and *target place of evacuation/displacement*. Then, we expanded the event extraction grammar, as described in section 4.4. In table 1 we list the clusters together with the main general and specific semantic category which they mostly represent, the corresponding prepositions with which these clusters were obtained and the event-specific semantic role, they were assigned.

5.2 *Experiments for Spanish*

For the Spanish language, we applied partially the learning algorithm described in section 4.3. We did not have time to collect necessary data for running the entire procedure. Instead of applying the whole algorithm, we translated two English-languages seed clusters into Spanish. More concretely, the first cluster contained four words, all designating different types of buildings and the second one consisted of three words, all designating disasters. Then, we applied the learning algorithm from step 4. That is, we performed cluster expansion using Ontopopulis and cluster

Table 1. Evaluation of the semantic consistency of the clusters (SF stands for Settlement and facility)

	size	gen. category	purity	sub-category	purity	prep.	ev.role
English							
c1	67	Calamity	85%	Natural disaster	67%	after	Cause
c2	48	Calamity	85%	Natural disaster	68%	after	Cause
c3	70	SF	70%	Facility	39%	to	Target
c4	95	SF	75%	Facility	58%	to	Target
c5	87	SF	71%	Facility	62%	into	Target
c6	69	Calamity	80%	Manmade disaster	56%	from	Cause
c7	111	SF	86%	Facility	84%	from	Source
c8	21	Situation	62%	Threat	33%	in	-
Spanish							
c9	72	Calamity	75%	Natural disaster	71%	-	Cause
c10	112	SF	55%	Facility	49%	-	-

Table 2. Accuracy of assigning event-specific roles using an extraction grammar

	Cause	Target place	Source (evacuated place)
English	86%	58%	100%
Spanish	32%	-	-

cleaning using Hearst patterns on the Web. In such a way, we obtained two extended semantic clusters for Spanish. In our experiments we used the extended cluster with the disasters to expand the Spanish event extraction grammar with one additional semantic role, namely *cause for displacement/evaluation*. We did not use the first stage of our algorithm which extracts seed terms (instead, the seed set was obtained as a translation of the English seed sets), therefore the clusters were not attached to specific prepositions. Regarding cluster $c10$, we did not include it in our grammar and therefore, we did not attach to it an event-specific role, however it can be used to find target or source places of evacuation and displacement events.

5.3 *Evaluation*

We carried out two types of evaluation: First, we evaluate the semantic consistency of each cluster and second, we run the extended event extraction grammar on a corpus of online news and extracted the entities,

which were assigned the newly added semantic roles; then we calculated the accuracy of assigning event-specific roles.

Semantic consistency was calculated by asking one English-speaking and one Spanish-speaking judge to define which is the semantic category which is predominant in each cluster. At first, the judge suggested quite generic categories, then they were asked to choose one more specific sub-category and to mark the cluster members which belong to the general category and to the more specific sub-category. Then, we calculated the *purity* of the cluster with respect to the generic and to the more specific categories as a ratio of the words which belong to the category and the cluster size. Results are presented in table 1.

The average purity of the English-language clusters with respect to the general category is 77%; it is 58% with respect to the more specific category. The corresponding purity values for the Spanish clusters is 65% and 60%. The purity of the Spanish-language clusters is comparable to the English ones. This is a good indicator for the multilingual nature of our algorithm. It is also important that cluster members, which we considered irrelevant for our evaluation, can still be considered relevant for the domain of displacements and evacuations: For example, our system learned words referring to vehicles and people, but they were mixed with other categories in the same cluster.

Regarding the extraction of event specific semantic roles, we run the extended grammar on an English and Spanish online news corpora, consisting of news clusters (each news cluster is a set of news articles, which refer to the same topic). For English we used a corpus of about 22,000 clusters and for Spanish we used a corpus of about 33,500 clusters. We calculated the accuracy of extraction for each of the event-specific semantic roles. We did not calculate recall, since at this stage our extended grammar was created mostly for experimental purposes and did not encode all the possible syntactic variations via which adjuncts can be connected to the event describing phrase. The results are presented in table 2.

The tangible result of our experiments was that new event specific roles were added to the event extraction grammar. In particular, the new slot $Cause$ was important, since it captured the events which lead to the displacement events.

The importance of detecting new semantic roles goes beyond extracting additional information. Detecting a semantic role, such $Cause$ together with some event-specific predicate, such as “flee” can be used to detect reliably an event of interest. For example, our grammar correctly

extracts "conflict" as a cause for displacement from the text: "...tens of thousands of civilians trying to $flee$ the $conflict$". Extracting such information allows us to detect reliably a report about a displacement event.

On the other hand, a word, such as "flee" alone does not provide enough evidence that an event of interest took place. For example, in the the following text "Meanwhile, the leopard injured several persons while making repeated attempts to $flee$.", no displacement event is described, still the word "flee" appears.

Most of the errors in our experiments were due to poorly clustered frequent words. For example, for Spanish we had the frequent words "pais" (country) wrongly clustered together with disasters. Similarly, the English word "killing" was clustered together with disasters and calamities, which lead to incorrect detection of a displacement event. With a little bit of manual cleaning, the accuracy of the obtained grammars could significantly be improved. Interestingly, some not very well classified words lead to grammar performance, which we considered correct. For example, the word "bomb" was clustered together with "war", "conflict" and other disastrous events. However, "bomb" is not an event. Nevertheless, the system extracts "bomb" as a cause for evacuation from the text: "Thousands of residents fled $bomb$-blasted parts of northern Mogadishu on Tuesday". This can be considered as a nearly correct match which lead to correct detection of an evacuation event, although strictly speaking the cause for evacuation was bombing and not "bomb". Similarly, "volcano" was clustered as a disaser and subsequently was extracted as cause for evacuation from the followiing text: "Thousands of people have been evacuated after a $volcano$ erupted" Such examples show that semantic similarities between words which belong to different categories (e.g. between "bomb" and "conflict") can be useful for practical purposes. Such kind of similarities cannot be found in semantic dictionaries, such as WordNet, however distributional word clustering successfully finds them. Clearly, distributional clustering is never 100% correct, however we think it is much easier to clean the errors from an already acquired dictionary, rather than creating one from scratch.

6 Conclusions and Future Work

In this paper we presented a multilingual algorithm for extending event extraction grammars by unsupervised learning of semantic classes. Although the results can be improved further, they show the viability of our approach.

The method we presented here can be used to automatize partially building of domain-specific grammars, which is quite a laborious task. As we demonstrated, the method can easily be adapted between languages.

Since our approach obtains word clusters, which model semantic concepts, it can also be used in the process of ontology building.

REFERENCES

1. Tanev, H., Piskorski, J., Atkinson, M.: Real-time news event extraction for global crisis monitoring. In: Proceedings of 13th International Conference on Applications of Natural Language to Information Systems, LNCS. (2008)
2. Faure, D., Nedellec, C.: A corpus-based conceptual clustering method for verb frames and ontology acquisition. In: LREC workshop on Adapting lexical and corpus resources to sublanguages and applications. (1998)
3. Li, H., Abe, N.: Generalizing case frames using a thesaurus and the mdl principle. Computational Linguistics **24** (1998) 214–244
4. Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of the Meeting of the Association for Computational Linguistics. (1990)
5. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the ACL'98. (1998)
6. Cimiano, P., Völker, J.: Towards large-scale, open-domain and ontology-based named entity classification. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria (2005)
7. Tanev, H., Magnini, B.: Weakly supervised approaches for ontology population. In: Ontology Learning and Population: Bridging the Gap between Text and Knowledge, IOS Press (2008)
8. Almuhareb, A., Poesio, M.: Extracting concept descriptions from the web: the importance of attributes and values. In: Ontology Learning and Population: Bridging the Gap between Text and Knowledge, IOS Press (2008)
9. Hearst, M.: Automatic discovery of wordnet relations. In: WordNet: An Electronic Lexical Database, MIT Press (1998) 131–152
10. Markert, K., Malvina, N., Modjeska, N.: Using the web for nominal anaphora resolution. In: EACL Workshop on the Computational Treatment of Anaphora. (2003) 39–46
11. Lian, S., Sun, J., Che, H.: Populating crab ontology using context-profile based approaches. In: Knowledge Science, Engineering and Management. (2007)

HRISTO TANEV
JOINT RESEARCH CENTRE, EUROPEAN COMMISSION,
VIA E. FERMI 2749, I-21027,ISPRA, ITALY
E-MAIL: <HRISTO.TANEV@EXT.JRC.EC.EUROPA.EU>

MIJAIL KABADJOV
JOINT RESEARCH CENTRE, EUROPEAN COMMISSION,
VIA E. FERMI 2749, I-21027,ISPRA, ITALY
E-MAIL: <MIJAIL.KABADJOV@JRC.EC.EUROPA.EU>

MONICA GEMO
JOINT RESEARCH CENTRE, EUROPEAN COMMISSION,
VIA E. FERMI 2749, I-21027,ISPRA, ITALY
E-MAIL: <MONICA.GEMO@JRC.EC.EUROPA.EU>