HIGH-COVERAGE EXTRACTION OF SEMANTIC ASSERTIONS FROM TEXT

Mitja Trampuš, Dunja Mladenić
Artificial Intelligence Laboratory, Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: mitja.trampus@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

We present an early version of a method for open-domain semantic assertion extraction from natural language texts. To combat the shortage of training data for the task, a two-stage pipeline is employed: we first perform semantic role labeling, then map the resulting frames onto predicate-form, ontology-aligned statements. We chose FrameNet and Cyc as the frame database and the ontology, respectively.

1 INTRODUCTION

In the majority of text mining tasks, algorithms operate at the syntactic level. Most often, the syntactic tokens are simply words. Such a representation is clearly limited in its expressive power: a lot of information is hidden in the interplay of the words. A standard method for accounting for this is the inclusion of word n-grams or other word co-occurrence structures into the data model. This approach, however, is not extensible indefinitely as it hits problems with sparsity.

A different approach to getting richer features from text is the use of parsers and the structural information they output. Thanks to advances in speed in accuracy (Charniak 2005; Collins 2003), this has been a particularly popular approach in the recent years. Because the approach has been proven advantageous in numerous applications, a natural next step is to increase the level of semanticity further and abstract the text into a purely logical form.

A big challenge in trying to perform such a mapping is the high cost of obtaining training data. Because data is further dependent on the choice of formalism, there is no large corpus which could be used for training. Data is, however, available for the task of semantic role labeling (SRL) (Toutanova 2007). At the same time, the FrameNet (Ruppenhofer 2008) collection of frames is quite semantic in nature, causing us to expect it can be mapped to an ontology reasonably well.

We therefore propose an approach for extracting ontology-based predicate assertions (in our case, Cyc) from plain text in two steps, using frame representation as a middle point. In this paper we describe both steps and show some results.

A quick note on notation: we use a sans serif typeface for sample sentences and *italics* for definitions of terms.

2 SEMANTIC ROLE LABELING

The task. Semantic role labeling (SRL) is a well-established text processing task in which the goal is to mark up text with a predefined set of *frames* and *frame elements,* also called *roles*. A *frame* is defined [Fillmore82] as any system of concepts (*roles*) related so that to understand any one concept it is necessary to understand the entire system.

Examples of frames are Addiction, Annoyance, Attack, Drinking etc. The latter, for instance, consists of roles Drinker, Fluid, Quantity, Container and perhaps others. There are also some roles that can be included in any frame, e.g. Location, Time, Frequency, Purpose and Manner. Not every occurrence of a frame in natural text needs fill all the roles; for example, the sentence [Paul $_{\text{DRINKER}}$] took a [sip $_{\text{TARGET}}$] of [red wine $_{\text{FLUID}}$] from [the tall glass $_{\text{CONTAINER}}$] and nodded approvingly. omits the Quantity role as well as all target-nonspecific roles. Note that this example uses the standard bracket notation for marking up frames. Also seen in the example is the [… $_{\text{TARGET}}$] role; this is a special role filled by the word that *evokes/triggers* the frame.

Another example of a frame would be BiologicalUrge: [He $_{\text{EXPERIENCER}}$] gave me a [tired $_{\text{TARGET}}$] [shrug $_{\text{EXPRESSOR}}$]. Frames are meant to be language independent.

Resources. There are two large resources available for training automatic SRL systems: FrameNet (Ruppenhofer 2008) and PropBank (Palmer, Kingsbury 2005). We describe and use FrameNet here, though many of he claims and approaches generalize to PropBank as well. FrameNet is a collection of frames and, importantly, frame-annotated sentences from various domains. There are 1020 frames, of which 540 have at least 40 annotated examples and 180 have at least 200. Each frame is also tagged with a list of trigger words (e.g. drink.v, drink.n, sip.v etc. for the Drinking frame). Every frame and every role is defined with a short natural-language definition. Frames are loosely connected with several relations, most notably generalization/specialization. For each pair of connected frames, the mapping between their roles is given as well.

The three stages of SRL. The process of automatic SRL decomposes naturally into three stages: frame identification (which frame is evoked by the sentence?), boundary detection (which sentence fragments are role fillers?) and role identification (what roles do the role fillers fill?). Although these problems can be solved jointly, it is easier and computationally much more efficient to approach them separately. This does not affect performance: it is intuitively clear that syntactic context should suffice for frame identification, but surprisingly, performing boundary detection and role identification jointly does not bring significant gains either (Gildea and Jurafsky 2002, Erk 2005). Our method thus performs each of the three stages separately as well.

Stage 1. For the frame identification task, we use a recall-oriented simplistic approach. First, we make the standard assumption that frames do not extend over more than one sentence. We then consider the lemmatized version of every word *w* in the sentence *s*. If, for any frame *f*, the lemma *w* occurs in *f*'s list of trigger words, we consider *s* to contain *f*. Some of these decisions are revoked at the later stages if no convincing role fillers are identified for *f* in *s*.

Stages 2 and 3. For role boundary detection, we first perform full constituency parsing of sentences using Charniak's (Charniak 2005) parser. We chose it over other parsers because it has state-of-the-art performance and is open-sourced which allowed us to modify it for online use. We then treat both remaining stages of SRL as classification tasks over the nodes of the parse tree.
We derive the following features for every node:

- Lemma of target word
- Phrase type (= Penn Treebank tag of node)
- Governing category (= parent node's tag; helps distinguish subjects from objects)
- Path from target to node
- Position relative to target (left/right)
- Passive/active voice of sentence. A sentence is considered passive if it contains an AUX↑VP↓VP↓VPN path.
- Lemma of node's head word. The head word is derived using widely adopted rules from Collins (1999).
- POS tag of node's head word.
- Verb subcategorization (= ordered list of children of VP immediately containing the node)

It has been shown that the choice of the classifier is not of critical importance; however, support vector machines (SVMs) are one of the most appropriate choices (Toutanova 2007; Carreras 2005). We use a linear SVM with ε=0.1, C=1/avg($\|x\|^2$) implemented in the svmlight toolset.
For stage 2 (role boundary detection) we use the above features and all of FrameNet's annotated data to classify each node as either *role* or *none*. We then discard all nodes which are classified as *none* with high confidence. The threshold was identified manually so that the pruning has about 95% recall and 55% precision. This significantly speeds up the role identification step and, perhaps even more importantly, greatly reduces class imbalance for the last step. In the role identification stage, we classify all the nodes remaining after the boundary detection stage into one of multiple classes: all the roles belonging to the frame and *none*. There is no clear consensus in the community on the best way to perform multi-class classification in this case, so we follow the recommendation from Hacioglu (2003) and use one-vs-all rather than pairwise classifiers or multi-class SVM.
When combining the votes, it is easy to satisfy the local constraints (each node should be assigned the class voted for with the highest confidence), but we should not neglect global constraints either (most importantly: a role appears only once in a frame, role fillers are strictly disjoint). We therefore employ a constrained greedy algorithm to assign roles. Votes for all nodes and all classes are sorted in descending order of confidence. They are then greedily assigned one by one; if an assignment would violate either of the two aforementioned global constraints, we discard the vote.
Additionally, based on observed algorithm bias towards nodes further from the root of the tree, we adjust the votes somewhat before sorting. Let us denote by $f(v,r)$ confidence of vote for role r on node v. If $f(v,r) > f(v, none)$ and, for some child node v' of v, it holds that $f(v',r) > f(v,r)$, then we set $f(v,r) := f(v',r)$.

Minor issues. To prepare training data, we map FrameNet's annotations (using word-level boundaries) onto parse tree nodes. In great majority of the cases, a perfect correspondence can be found; if, due to errors in parsing or due to a convoluted sentence structure, a perfect match does not exist, we map the role-filler annotation to the leftmost highest node in the tree which is completely contained the in annotation. We noticed that in English, this tends to preserve the semantic head of the role filler.
Akin to most of the existing work, we build a separate set of classifiers for every frame. This could be improved by taking into account that some roles (e.g. Place, Time) are shared across frames.
We limit ourselves to frames that describe actions, e.g. Drinking but not BiologicalState. There are several reasons for this: action frames are more informative, map to Cyc more cleanly and have better annotation coverage in FrameNet. Action frames were identified by having at least one verb trigger word and not more than 10 times as many non-verb trigger words. Of those, we discard frames with no annotated sentences. By hand inspection, we discarded further 20 frames deemed too generic or irrelevant (e.g. Undergoing with the definition "An Entity is affected by an Event."). We are left with approximately 550 frames. We also considered using roughly 100 additional frames where one of the roles generalizes to the generic Actor or Experiencer role, but decided against it for now as their mapping to Cyc is less straightforward (they mostly do not correspond to an #$Event; see section 3.1).

3 MAPPING FRAMENET TO CYC

As discussed in the introduction, our end goal is to obtain a semantic representation of input text. The SRL markup obtained using the method from the previous section, though, marks up syntactic constituents of the sentence. We thus still need to map the role fillers to an ontology. In general, this task is no easier than the one we started out with (mapping whole sentences), because role fillers can be whole relative clauses: for example, for frame Drinking, we can have the sentence [He $_{DRINKER}$] [drank $_{TARGET}$] [the strange stink emitting potion she had concocted for him before they left for the journey $_{FLUID}$]. Mapping the Fluid role onto a set of ontological concepts is clearly no different from the original task. Luckily, it is reasonable to assume that the extra properties about the potion will be identified during analysis of other frames, e.g. Cooking: He drank [the strange stink emitting potion $_{FOOD}$] [she $_{COOK}$] had [concocted $_{TARGET}$] [for him $_{PURPOSE}$] [before they left for the journey $_{TIME}$]. and Apperance[1]: He drank the strange [stink $_{TARGET}$] emitting [potion $_{PHENOMENON}$] she had concocted for him before they left for the journey.

Our problem therefore reduces to mapping only the semantic head of each role filler. The head is either a noun phrase (potion) or a verb phrase (meet in [He $_{EXPERIENCER}$] [hoped $_{TARGET(DESIRING)}$] [to meet her again $_{EVENT}$].) or a noun phrase. The second case is easy to resolve: verb phrases are almost without exception targets of frames themselves; a verb-phrase head is therefore mapped simply to a whole frame (in the above example, SocialEvent). In the case of noun-phrase heads, we choose to simplify by discarding all adjectival information. This is motivated similarly to our limitation on a subset of frames: information conveyed by adjectives is in general less crucial and has poorer support in ontologies. What remains to be mapped is a very short noun phrase, typically consisting of a single word; in other words, we are left with the task of word sense disambiguation (WSD).

Before mapping the role filler, we of course have to choose an ontology. Ideally, this would be FrameNet as our frames and roles already come from it. However, FrameNet is not a general-purpose ontology; it does cover some entities (presented as frames that tend to require no roles), but for example has no satisfactory mapping for dog or seat. We therefore chose Cyc, an ontology created specifically for purposes like this and containing millions of concepts related to common knowledge.

The choice of Cyc (or any other ontology different from FrameNet) however introduces the necessity to map the frames and roles as well. This problem is known as ontology alignment.

We next describe our approach to both tasks.

[1] A somewhat unfortunately generic frame; taken, however, verbatim from FrameNet

3.1 Mapping Frames and Roles

Conceptually, it makes sense to first align the ontologies for two reasons. First, this is a task that only needs to be done once. Second, it offers support for WSD in that the ontology imposes selectional preferences and constraints on role fillers using its type system. This can aid in the role identification phase of SRL or at least be used immediately after it in a reranking postprocessing step. Our approach currently does not yet make use of this.

(Dis)similarities between the ontologies. Of the numerous concepts found in Cyc, of special interest to us are #$Event and #$BinaryRolePredicate. Specializations of the first are a natural counterpart of FrameNet's frames. Instances of the second are the counterpart of FrameNet's roles. They are connected by the #$rolesForEventType relation which specifies which roles apply to which events. In short, the structure of that part of Cyc is quite similar to that of FrameNet[2]. A majority of frames has a natural counterpart that is a specialization of the #$Event concept in Cyc. We currently discard the frames that do not; those fall in one of the following categories:

- Frame maps to more than one Cyc concept. For example, the frame Respond_to_proposal (with triggers reject, accept, refuse etc.) could map to Cyc's #$Refusing-CommunicationAct, #$Accepting-CommunicationAct, #$Rejecting-CommunicationAct and some others, but their only common generalization is #$CommunicationAct, which is too general. About 5% of frames are like this.
- Concept does not exist in Cyc. For example, Adjusting (triggers: adjust, tweak, calibrate, …). This does not necessarily mean the notion cannot be expressed in Cyc, but it would require a non-atomic expression. About 2% of frames fall into this category.
- About 2% of the frames map to relations rather than specializations of #$Event. For example, Evoking maps to the relation (#$evokes ARG1 ARG2) where ARG1 is an instance of #$Individual and ARG2 of #$FeelingAttribute.

With a moderate amount of additional work, frames from the last two categories could be accommodated as well, meaning that 95% of the frames we consider have a natural counterpart in Cyc. This supports our decision to use FrameNet for an intermediary representation of information.

It has to be noted, however, that not all mappings are perfect. In particular, we are sometimes forced to ignore certain subtleties in frame definitions. Consequently, several FrameNet frames might get mapped to the same Cyc concept. An extreme example of this is the #$Evaluating concept which is mapped to by Trying_out, Labeling,

[2] and it would be very reasonable to perform SRL directly using Cyc as the frame ontology, were it not for a complete lack of training data.

Regard, Judgment and Assessing. Another typical example of conflated frames are frame pairs of the form Cause_to_XYZ and XYZ. We map pairs like this to the same Cyc concept, but with different role mappings.

Semi-supervised mapping of frames. There are about 550 frames to be mapped and about 2000 roles. While the best accuracy would certainly be achieved by mapping by hand, this is prohibitively time-consuming. On the other hand, automatic approaches have few reliable features and no training data, so a completely unsupervised approach is also unrealistic. We opt for a semi-supervised scenario where an algorithm proposes several possible mappings and a human annotator chooses the best one among them.

When aligning ontologies, there are, broadly speaking, two types of features available: content-based, stemming from the properties of the nodes themselves (typically, glosses or sample instances), and structural. In our case, aiming at aligning the two ontologies structurally does not make sense as the two have different levels of granularity and coverage. We therefore make use only of the glosses and English denotation strings of entities in both ontologies.

When mapping frames, the trigger words provided with each frame prove to be much more valuable than the frame descriptions. Our method suggests for each frame all the concepts that have at least one of the trigger words of the frame listed as their English denotation. It also suggests all the common ancestors of these initially collected Cyc concepts in the generalization taxonomy: for example, the frame Inchoative_change_of_temperature is associated, among others, with trigger words chill, cool and heat. In Cyc, cool is not associated with any concept (English annotations are lacking), chill is associated with #$Chilling and heat is associated with #$HeatingProcess. One of their common ancestors is #$TemperatureChangingProcess, which is the right mapping for the frame in question.

To ease the annotator's job, the suggested Cyc concepts are ranked according to the number of FrameNet trigger words that map to them and their depth in Cyc taxonomy (more specific is better). However, the ranking did not prove essential as the number of suggestions is typically low. We therefore did not experiment with more complex ranking approaches based e.g. on similarities of glosses.

Automatic mapping of roles. Even with the semi-supervised approach, the time investment for mapping roles would be too large given their number. We therefore perform the mapping automatically, based on heuristics only. To increase accuracy, we only map the core roles[3] of each frame. This corresponds to roughly 80% of roles appearing in natural text. In Cyc, we do not have such information and therefore consider all roles; however, we discard those for which a more specific role (according to role hierarchy) is available as well.

To determine role similarity, we use the glosses and subject/object information. From glosses, a bag of words vector is constructed (with tf-idf weighting, Porter stemming and a stopword list). By subject/object information, we mean that the two most important roles tend strongly to be the subject and the object. For all Cyc roles, it is possible to infer (using role hierarchy) what the subject and the object are, if any. For FrameNet roles, a similar inference is sometimes possible (the hierarchy is much less principled and populated); when hierarchical info is unavailable, we heuristically assume that the first role to be listed is the subject with probability 0.7 and object with probability 0.3; and the other way around for the second role listed. For roles that have been identified as subjects or objects, this is added as an extra component to the sparse vector.

We define role similarity as the cosine between the two length-normalized vectors. To obtain the best global assignment, we create a bipartite graph of roles and weigh every edge connecting two roles r and r' with

$$w(r,r') := d(r,r')^{0.5}$$

where d is the cosine similarity between the feature vectors. We then use Hungarian method to find the maximum-weight assignment. The square root was introduced to further decrease the "greediness" of the method (propensity to choose the highest-scoring pair regardless of others). Another possible regularization is logarithmic (treating similarity scores as probabilities; the probability of the global assignment is then the product of pairwise probabilities, i.e. the sum of logarithms). We have also experimented with a few naïve greedy approaches, but found their performance to be worse.

In the above approach, we assume that no two roles from FrameNet map onto a single role in Cyc. It should be noted that this can be problematic. Especially for actions with "symmetric" roles, FrameNet assumes a somewhat confusing notation: for example, the frame Meeting contains roles Party_1, Party_2 and Parties. Some frame occurrences fill the first two roles and others fill only the third role – depending on the phrasing. In Cyc, all of these correspond to a single role (which may then have two distinct fillers).

3.2 Mapping Role-Fillers (WSD)

Identifying the head. For a role filler, we first identify its semantic head. This is different from the syntactic head used in the feature construction stage of SRL, so we derive a separate set of simple recursive rules. For NP nodes, descend into the rightmost noun-like child. For PP nodes that start with a preposition, descend into the child immediately following it. For S nodes, descend into the last verb (phrase). For VP nodes descend into the first verb (phrase). If no rule applies, stop.

Choosing among mappings. For mapping role fillers to Cyc, we use Cyc's built-in #$termStrings predicate which

[3] This is a FrameNet concept. Core roles are those that either have to be appear explicitly or their filler is implicitly understood from the context. A frame typically has two to four core roles.

connects concepts and English words. Often, a single word maps onto multiple concepts. At the task of WSD, simply mapping to the most common interpretation for the word will give an extremely strong baseline. Unlike WordNet, Cyc unfortunately has no “most common sense” information associated with each word. It does, however, have links from its concepts to WordNet. Although created semi-automatically and not of perfect quality or coverage, they allow us to rank all the Cyc concepts suggested by #$termStrings using commonness information from their WordNet counterparts. The highest ranking concept is then selected. If there are multiple highest-ranking concepts or if there is no WordNet information available due to absence of links, we give priority to the concepts first returned by the Cyc inference engine.

This is a very simple approach; we plan to later integrate a separate Cyc WSD engine currently being developed at our department.

4 RESULTS

SRL. To make our system more comparable with existing ones, we only measure performance on 10 frames, training on 300 annotated sentences. We achieve precision 56% and recall 61%. While we acknowledge that these results are lower than the state of the art (F1 in high seventies (Litkowski 2004)), there is also clear room for improvement. We expect our decision not to use out-of-the-box SRL packages to prove beneficial when we improve the pipeline as a whole and increase coupling between the SRL and the ontology alignment phase.

As described, we do not perform frame identification beyond trigger keyword matching, so we cannot comment on its performance.

Framenet-Cyc alignment. In this step, it makes little sense to compare ourselves with existing contributions to the field as the achievable performance depends highly on the actual ontologies we are trying to align.

For frame alignment (without the roles), we used a human annotator as described in section 3.1. There was only one annotator, so inter-annotator agreement has not been measured.

To evaluate the role alignment step, we manually inspected all 83 core roles in 25 randomly selected frames successfully mapped to Cyc. Accuracy is 35/83=42%; a perfect mapping could achieve at most 64/83=77% on this sample since for some frames, the corresponding Cyc concept is not associated with enough roles. We do have to note that mapping accuracy on the subject and object roles is higher, and because real-world sentences use these two roles more than others, the error rate introduced will be somewhat better than what the 42% above suggest.

Role filler alignment (WSD). Based on manual inspection of 50 role fillers, we estimate that the semantic head of the role filler is correctly identified in 78% of the cases. Mapping of role fillers to Cyc is correct in 60% of the cases in which the semantic head is identified correctly and thus in 48% of the cases overall. In this count, we ignore the pronouns he, she, her, him and his which are mapped to the generic #$Person concept or its gender specializations with hand-written rules.

Overall performance. It is very hard to estimate the recall of the complete pipeline or indeed even of SRL alone as there is no strict enough definition of what a frame is. If, however, for the sake of evaluation, we assume that FrameNet has perfect coverage, the recall of the pipeline at the frame level (i.e. frames successfully identified and mapped onto Cyc with at least one role) is about 65%. For intra-frame performance, refer to the previous paragraph.

As an illustrative example, we are attaching an excerpt from a newspaper article along with the automatically extracted facts. The text:

To understand and appreciate the Bush administration's policy regarding Israeli Prime Minister Sharon's disengagement plan, we must briefly reexamine the record. For three and a half years now, the administration's attitude toward the Israeli-Palestinian conflict/peace process has been characterized by high rhetoric and little action. On the one hand, President Bush is the first US leader to officially endorse the creation of a Palestinian state.

Facts from the first sentence:

(#$objectImproved #$Comprehending* #$OrganizationPolicy*)
(#$performedBy #$Comprehending* (ObjectDenotedByFn "we"))
(#$evaluationInput #$Evaluating* #$OrganizationPolicy*)
(#$performedBy #$ExercisingAuthoritativeControlOverSomething* (ObjectDenotedByFn "we"))
(#$performedBy #$PurposefulAction* (ObjectDenotedByFn "Sharon"))

Facts from the second sentence:

(#$eventOccursAt #$DescribingSomething* #$Attitude*)
(#$senderOfInfo #$DescribingSomething* #$Action*)
(#$performedBy #$ExercisingAuthoritativeControlOverSomething* (ObjectDenotedByFn "constitutes"))

Facts from the third sentence:

(#$performedBy #$Siding-SelectingSomething #$Bush*)
(#$doneBy ArrivingAtAPlace #$Bush)
(#$communicatorOfInfo #$Communicating #$Bush)

Some facts are very sensibly extracted (the first sentence does particularly well) while some of them are highly erroneous (e.g. most from the second sentence, or president Bush being mapped to #$Bush, the garden bush concept). A word on notation: With a star, we denote here “an instance of collection”: for example, #$Dog is specified in Cyc to denote the collection of all dogs, so we use #$Dog* in the example above to denote a specific instance of #$Dog. In actual program output, this is denoted with multiple predicate statements and using #$isa. The notation (#$ObjectDenotedByFn “foo”) means a concept Cyc does not know about, but is expressed in English as “foo”.

5 RELATED WORK

SRL methods are well researched and numerous. Their basic design is unchanged since the first reported attempt at SRL (Jurafsky 2003). A basic preprocessing step is constituency parsing (although a few rare examples opt for chunking or other shallower methods (Punyakanok 2004). This gives rise to most of the features; feature engineering was shown to be very important (Toutanova 2006). The problem is then typically divided into role detection and role identification steps; both are almost always performed using classic ML techniques. The best insight into SRL is offered by various challenges (Litkowski 2004; Carreras 2005; Ruppenhofer 2010).

The task of semantic fact extraction is much less researched. The better-known systems aim for high precision; this means that they only search for a limited number of relations and even within those do not focus on recall. TextRunner (Banko 2008) is is an example of such a system, though it does not completely meet our criteria in that the relations and entites it extracts are stil represented as textual strings. SOFIE (Suchanek 2009) is a recent system that performs ontology alignment as well. High-recall, general-domain oriented fact extraction has been attempted by Rusu (2009) by focusing on subject-verb-object triplets. The output is textual. Role filler alignment corresponds largely to the task known as Word Sense Disambiguation; refer to Navigli (2009) for a recent survey.

6 CONCLUSIONS AND FUTURE WORK

As demonstrated by the evaluation, automated fact extraction still has a long way to go. Because of the large gap between the textual and purely semantic representation, it is almost inevitable for approaches to employ long pipelines. While it is possible to achieve reasonable accuracy at each individual step, the pipeline length means a large number of errors accumulates. We believe this would remain a problematic factor even if we improved our individual methods – for which there is ample room. The solution is most likely in merging pipeline stages; in our approach, for example, the role labeling and Cyc mapping could be done in a more intertwined fashion, allowing the two to correct each other.

In the future, we would like to explore SRL based on simpler structural features, e.g. chunker output. There are several motivating factors for that. First, in the context of domain independence, full-parse features are problematic (Huang 2010, Croce 2010) because parsers are typically trained on the Penn Treebank (= annotated Wall Street Journal articles) and do not generalize well to other domains; SRL, in turn, shows high dependence on parser accuracy. Second, full constituency parsing is still quite slow and third, they are available for fewer languages. Last but not least, the simplification in features seems to affect performance by only about 2% (Surdeanu 2007).

Also related to SRL, we would like to explore ways of automatically increasing the amount of training data. We still see the lack of data as a major impediment; most papers and challenges on SRL limit themselves to only the few best-annotated frames.

As the primary motivation for extraction of semantic assertions is their further potential utility in text mining tasks, we plan to test their usefulness in this manner as well. In particular, because we chose an ontology with relatively large amounts of background knowledge and good inferencing capabilities, we would like to explore the value of facts inferred from the ontology.

7 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and RENDER (ICT-257790-STREP).

References

Banko, M., Etzioni, O., Center, T.:: The tradeoffs between open and traditional relation extraction. In: Proceedings of ACL 08, pp. 28–36. Citeseer, (2008)

Carreras, X., Màrquez, L.:: Introduction to the CoNLL-2005 shared task: Semantic role labeling. In: Proceedings of the Ninth Conference on Computational Natural Language Learning, pp. 152–164. Association for Computational Linguistics, (2005)

Charniak, E., Johnson, M.:: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 173–180. Association for Computational Linguistics, (2005)

Collins, M.:: Head-driven statistical models for natural language parsing, PhD-thesis, University of Pennsylvania (1999)

Collins, M.:: Head-driven statistical models for natural language parsing. Computational linguistics 29, 589–637 (2003)

Croce, D., Giannone, C., Annesi, P., Basili, R.:: Towards Open-Domain Semantic Role Labeling. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 237–246. Association for Computational Linguistics, (2010)

Hacioglu, K., Ward, W.:: Target word detection and semantic role chunking using support vector machines. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology companion 25-27 (2003)

Huang, F., Yates, A.:: Open-domain semantic role labeling by modeling word spans. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 968–978. Association for Computational Linguistics, (2010)

Litkowski, K.:: Senseval-3 task: Automatic labeling of semantic roles. In: Senseval-3: Third International Workshop on the, pp. 2-5. (2004)

Navigli, R.:: Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) (2009)

Palmer, M., Gildea, D., Kingsbury, P.:: The proposition bank: An annotated corpus of semantic roles. Computational Linguistics 31, 71–106 (2005)

Punyakanok, V., Roth, D., Yih, W.-tau.:: The necessity of syntactic parsing for semantic role labeling. In: International Joint Conference on Artificial Intelligence, pp. 1117. Citeseer, (2005)

Punyakanok, V., Roth, D., Yih, W., Zimak, D.:: Semantic role labeling via integer linear programming inference. In: Proceedings of the 20th international conference on Computational Linguistics, pp. 1346–es. Association for Computational Linguistics, (2004)

Ruppenhofer, J. et al.:: FrameNet II: Extended Theory and Practice (2006)

Ruppenhofer, J. et al.:: Semeval-2010 task 10: Linking events and their participants in discourse. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 45–50. Association for Computational Linguistics, (2010)

Rusu, D., Fortuna, B., Grobelnik, M., Mladenic, D.:: Semantic Graphs Derived From Triplets With Application In Document Summarization. Informatica Journal (2009)

Suchanek, F.M., Sozio, M., Weikum, G.: SOFIE: a self-organizing framework for information extraction. In:Proc. of the 18th international conference on World wide web, pp. 631–640. ACM, (2009)

Surdeanu, M., Màrquez, L.: Combination strategies for semantic role labeling. Journal of Artificial Intelligence (2007)

Yih, S.W., Toutanova, K.:: Automatic semantic role labeling. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts on XX, pp. 309–310. Association for Computational Linguistics, (2006).