Open Domain Information Extraction via Automatic Semantic Labeling

Alessandro Moschitti
Department of Computer Science
Systems and Production
University of Rome Tor Vergata
00133 Rome (Italy)
moschitti@info.uniroma2.it

Paul Morărescu
Department of Computer Science
University of Texas at Dallas
Richardson, TX 75083-0688
paulm@student.utdallas.edu

Sanda M. Harabagiu
Department of Computer Science
University of Texas at Dallas
Richardson, TX 75083-0688
sanda@utdallas.edu

Abstract

This paper presents a semantic labeling technique based on information encoded in FrameNet. Sentences labeled for frames relevant to any new Information Extraction domain enable the automatic acquisition of extraction rules for the new domain. The experimental results show that both the semantic labeling and the extraction rules enabled by the labels are generated automatically with a high precision.

Keywords: Natural Language Processing, Information Extraction.

Introduction

With the advent of the Internet, more and more information is available electronically. Most of the time, information on the Internet is unstructured, generated in textual form. One way of automatically identifying information of interest from the vast Internet resources is by employing Information Extraction (IE) techniques.

IE is typically performed in three stages. First, the information need is abstracted and expressed as a structured set of inter-related categories. These structures are called templates and the categories that need to be filled with information are called slots. For example, if we want to extract information about natural disasters, we may be interested in the type of disaster, the damage produced by the disaster, in the casualties as well as in the date and location where the disasters occurred. Therefore, we may generate a template listing such categories as *DAMAGE*, *NUMBER_DEAD*, *NUMBER_INJURED*, *LOCATION* and *DATE*.

Second, as the extraction template is known, text snippets containing the information that may fill the template slots need to be identified. The recognition of textual information of interest results from pattern matching against extraction rules, which are very much dependent on the knowledge of the domain of interest. For example, if we want to extract information about natural disasters, we need to recognize (a) types of disasters, names of locations and dates; and (b) all the syntactic alternations of expressions that report to natural disasters, e.g.:

`"A tornado hit Dallas Monday at 8am."` or
`"Reports on a tornado touch down in Dallas came as early as 8 in the morning."` or
`"Two people were injured when a tornado touched down in Dallas last Monday."`

In the third phase, after information of interest is identified in the text of electronic documents, it needs to be mapped in the correct template slot. This mapping is not trivial, as rarely we can identify in the same sentence all fillers of a template.

All these phases of IE are dependent on knowledge about the events, states or entities that are of interest, also known as *domain knowledge*. Every time when the information of interest changes, new domain knowledge needs to be acquired and modeled in the extraction rules. This task is complex, as it has been reported in (Riloff & Jones 1999; Harabagiu & Maiorano 2000; Yangarber *et al.* 2000), and it requires both high quality seed examples and texts relevant to the extraction domain. The two limitations hinder the extension of IE techniques to virtually any topic of interest, or Open-Domain IE.

The recent availability of the FrameNet lexico-semantic database (www.icsi.berkeley.edu/∼framenet) allows us to reconsider the problem of Open-Domain Information Extraction. The aim of the FrameNet project is to produce descriptions of words based on semantic frames. Semantic frames, as they have been introduced by (Fillmore 1982), are schematic representations of situations involving various participants, properties and roles, in which a word may be typically used. This kind of knowledge can be successfully used for generating domain knowledge required for any new domain, i.e. Open-Domain Information Extraction. The corpus annotation available from FrameNet enable us to generate a labeling procedure that allows the recognition of extraction rules for any domain.

The remainder of this paper is organized as follows. Section 2 details the FrameNet data and contrasts it against extraction templates whereas Section 3 shows our method of learning semantic frame categorization by employing Support Vector Machines (SVMs). Section 4 presents our experiments in Open-Domain IE and

Section 5 summarizes our conclusions.

Semantic Frames

The Semantic Frames available from FrameNet are in some way similar to efforts made to describe the argument structures of lexical items in terms of case-roles or thematic-roles. However, in FrameNet, the role names, which are called Frame Elements (FEs) are local to particular frame structures. Some of these FEs are quite general, e.g. *AGENT*, *PHENOMENON*, *PURPOSE* or *REASON*, while others are specific to a small family of lexical items, e.g. *EXPERIENCER* for *Emotion* words or *INTERLOCUTOR* for COMMUNICATION words. Most of the frames have a combination of FEs, some are general, some are specific. For example, the FEs of the ARRIVING frame are *THEME*, *SOURCE*, *GOAL* and *DATE*. They are defined in the following way: the *THEME* represents the object which moves; the *SOURCE* is the starting point of the motion; the *PATH* is a description of the motion trajectory which is neither a *SOURCE* nor a *GOAL*; the *GOAL* is the expression which tells where the theme ends up.

A frame has also a description that defines the relations holding between its FEs, which is called the *scene* of the frame. For example, the scene of ARRIVING is: the *THEME* moves in the direction of the *GOAL*, starting at the *SOURCE* along a *PATH*. Additionally, FrameNet contains annotations in the British National Corpus (BNC) of examples of words that evoke each of the frames. Such words are called *target words*, and they may be nouns, verbs or adjectives. Although all these three major lexical categories can be frame bearing, the most prominent semantic frame evoked in a particular sentence is usually one evoked by a verb. For example, the target words evoking the ARRIVING frame are: approach(v), arrival(v), arrive(v), come(v), enter(v), entrance(n), return(n), return(v), visit(n) and visit(v) [1].

S1: [Yorke]$^{\text{FE=THEME}}_{\text{PT=NP GF=Ext}}$ [returning]$_{\text{TARGET}}$ [home]$^{\text{FE=GOAL}}_{\text{PT=AVP GF=Comp}}$
[from a charity event]$^{\text{FE=SOURCE}}_{\text{PT=PP GF=Comp}}$ at 2am, the city's magistrates heard.

S2: [Returning]$_{\text{TARGET}}$ [across the square]$^{\text{FE=PATH}}_{\text{PT=PP GF=Comp}}$ [she]$^{\text{FE=THEME}}_{\text{PT=NP GF=Ext}}$
felt she was going home; not for one moment did she confuse such a place with the Aber House Hotel.

S3: You heard [she]$^{\text{FE=THEME}}_{\text{PT=NP GF=Ext}}$ [returned]$_{\text{TARGET}}$ [heartlessly]$^{\text{FE=MANNER}}_{\text{PT=AVP GF=Comp}}$.

Figure 1: Example of sentences mapped in FrameNet

In FrameNet the annotations seek to exemplify the whole range of syntactic and semantic dependencies that the target word exhibit with any possible filler of a FE. For example, Figure 1 shows four FrameNet annotations corresponding to the verb *return*. The FrameNet tagset used to annotate the BNC sentences contain different tags which were described in (Johnson & Fillmore 2000). In our experiments we relied only on these tags: (1) the *target word* (*TARGET*); (2) the *phrase type* (PT); and (3) the grammatical function (GF). The first sentence illustrated in Figure 1 has annotations for the *THEME*, *GOAL* and *SOURCE* FEs, whereas the second sentence has an annotation for the *PATH* frame element. The annotations from Figure 1 also use different possible values from the phrase type (PT) tags and the grammatical function (GF) tag. These values are listed in Tables 1 and 2. Sentence S3 contains an annotation for *MANNER*. Figure 2 illustrates a part of the FrameNet hierarchy. Sometimes multiple frames have the same FEs, e.g. the ARRIVING and DEPARTING frames, but their *scenes* contrast their semantic interpretation.

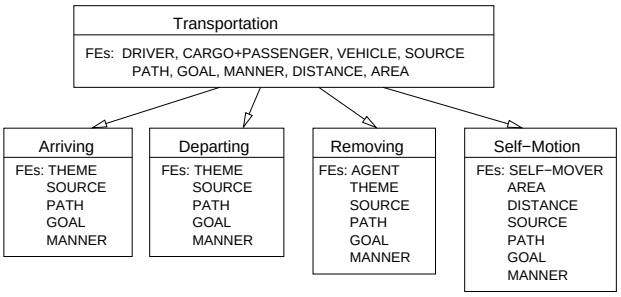

Figure 2: Hierarchical structuring of the Motion domain in FrameNet

Table 1: Phrase types annotated in FrameNet

Label	Phrase Type Description
NP	Noun Phrase (*the witness)*
N	Non-maximal nominal (personal *chat*)
Poss	Possessive NP (*the child's* decision
There	Expletive *there* (*there* was a fight)
It	Expletive *it* (*it*'s nice that you came)
PP	Prepositional phrase (look at *me*)
Ping	PP with gerundive object (keep *from laughing*)
Part	Particle (look it *up*)
VPfin	Finite verb phrase (we *ate fish*)
VPbrst	Bare stem VP (let us *eat fish*)
VPto	To-marked infinitive VP (we want to *eat fish*)
VPwh	WH-VP (we know *how to win*)
VPing	Gerundive VP (we like *winning*)
Sfin	Finite clause (it's nice *that you came*)
Swh	WH-clause (ask *who won*)
Sif	*If/whether* clause (ask *if we won*)
Sing	ve clause (we saw *them running*)
Sto	To-marked clause (we want *them to win*)
Sforto	*For-to* marked clause (we would like *for them to win*)
Sbrst	Bare stem clause (we insist *that they win*)

The FrameNet structures and their annotations can be used for extracting information in a topic that relates to the domains they encode. To experiment with the usage of FrameNet for IE, we have employed the extraction definitions used in the Hub-4 Event'99 evaluations (Hirschman *et al.* 1999). The purpose of this extraction

[1] n stands for noun and v stands for verb.

task was to capture information on certain newsworthy classes of events, e.g. natural disasters, deaths, bombings, elections, financial fluctuations. Extraction tasks do not use frames, but instead they produce results in the form of templates. For example, let us consider the template devised for capturing the movement of people from one location to another. Individual templates were generated for fifteen different generic events.

Table 2: Grammatical functions annotated in FrameNet

Label	Grammatical Function Description
Ext	*External argument* (Argument outside phrase headed by target verb, adjective or noun)
Comp	*Complement* (Argument inside phrase headed by target verb, adjective or noun)
Mod	*Modifier* (Non-argument expressing FE of target verb, adj. or noun)
Xtrap	*Extraposed* (Verbal or clausal compl. extraposed to the end of VP)
Obj	*Object* (Post-verbal argument; passivizable or not alternate with PP)
Pred	*Predicate* (Secondary predicate compl. of target verb or adjective)
Head	*Head* (Head nominal in attributive use of target adjective)
Gen	*Genitive determiner* (Genitive determ. of nominal headed by target)

We have used these templates for studying ways of mapping their slots into FEs of FrameNet frames. We have noticed that one Event'99 template is generally mapped into multiple FrameNet frames. The slots of the template are: *PERSON*, *FROM_LOCATION*, *TO_LOCATION* and *DATE*. Figure 3 illustrates a mapping from the slots of this template to the FEs of two different frames encoded in FrameNet.
In our experiments we have manually produced the mappings. Since mappings are possible from any given template to FEs encoded in FrameNet, we developed a five-step procedure of acquiring domain information in the form of extraction rules for any topic. The procedure is:

Open-domain Information Extraction (Template)

1. *Map Template slots into the FEs of frames from FrameNet.*
2. *Given a text, label each sentence either with F_A, if it contains information from the domain of frame A, or with ϕ.*
3. *In each labeled sentence identify:*
   - *3.1 the target word*
   - *3.2 instantiations of FEs from frame A*
4. *For each verb identified as*
   - *(a) target word or in a Subject-Verb-Object dependency with the target word; or*
   - *(b) in a FE instantiation*

   collect all Subject-Verb-Object triplets as well as all the prepositional attachments of the verb;
5. *Generate extraction rules for the topic.*

The result of this procedure is that we obtain as many extraction rules as many different verbs we have identified. Their subjects, objects and prepositional objects are matched by any nouns groups having the head in the same semantic category as those learned at training time from the FrameNet annotations. Central to this procedure is step 2, which identifies relevant sentences. Based on this categorization, we can perform step 3 with high-precision, in a second labeling pass.

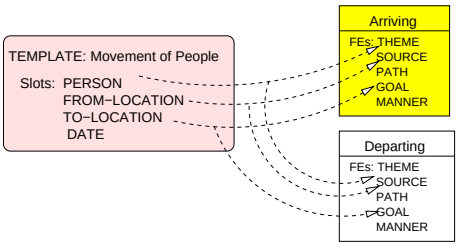

Figure 3: Mappings from an extraction template to multiple frames.

Semantic Labeling

The first pass of labeling concerns identifying whether a sentence contains information pertaining to a frame encoded in FrameNet or not. It is possible that a sentence is relevant to two or multiple frames, thus it will have two or multiple labels. In the second pass text snippets containing a target word and the instantiation of a frame elements are detected.

Sentence labeling

The problem of semantic labeling of sentences is cast as a classification problem that can be trained on the BNC sentences annotated in FrameNet.

To implement the classifier we have chosen the Support Vector Machine (SVM) model because it is known it generally obtains high classification accuracy without requiring high quality training data (Vapnik 1995). In our SVM-based procedure we have considered the following set of features: each distinct word from the training set represents a distinct feature; additionally, each distinct <*Phrase Type - Grammatical function*> pair (<*PT-GT*>) that is annotated in the training set represents a distinct feature. In our experiments, we have used 14,529 sentences containing 31,471 unique words and 53 distinct <*PT-GF*> pairs. The total number of features was N=31,524. The sentences were selected from the FrameNet examples corresponding to 77 frames.

For each frame F_a we have trained a different classifier C_a. Considering each sentence s from the training corpus T_c, we generate its feature wegiht vector $\vec{f_s} =< f_1^s, f_2^s, ..., f_N^s >$. The value of $f_i^s = 1$ only if the i-th feature was observed in sentence s, otherwise $f_i^s = 0$.

We have also computed the weight for each feature f_i^s observed in the sentence. First we have computed the feature's frequency in the sentence s. Then, we have measured N_{f_i}, the number of sentences of the training corpus T_c that contain f_i. We have obtained the *Inverse Sentence Frequency* for feature f_i as $ISF(f_i) = log(\frac{n_c}{N_{f_i}})$, where n_c is the total number of sentences in T_c. The weight for a single feature is given by:

$$\omega_{f_i}^s = \frac{r_{f_i}^s \cdot ISF(f_i)}{\sqrt{\sum_{j:f_j{}^s=1}[r_{f_j}^s \cdot ISF(f_j)]^2}}$$

The weighted feature vector associated to a sentences s will be $\vec{w_s} =< \omega^s_{f_1}, \omega^s_{f_2}, ..., \omega^s_{f_N} >$.

To classify a new sentence s' for a frame F_a we need to learn a linear function $l_a(\vec{x}) = \vec{a} \cdot \vec{x} + b$ in which $\vec{x}$ is the feature vector for s' at classification time. The values of vector $\vec{a}$ and of b are obtained from the resolution of the following optimization problem:

$$\begin{cases} Min \quad ||\vec{a}|| \\ \vec{a} \cdot \vec{w_s} + b \geq 1 \quad \forall s \in T_c \ \ labeled \ \ for \ \ F_a \\ \vec{a} \cdot \vec{w_s} + b \leq -1 \ \ \forall s \in T_c \ \ not \ \ labeled \ \ for \ \ F_a \end{cases}$$

The SVM classifier C_a for the frame F_a applies the signum function (sgn) to the linear function l_a, i.e. $C_a(\vec{x}) = sgn(l_a(\vec{x}))$. A sentence s' is labeled for F_a only if $C_a(\vec{w_{s'}}) = 1$. In our experiments, we have used the SVM implementation from the Rainbow package (McCallum 1996).

Refining Semantic Labels

For the purpose of open-domain IE, we need to know additionally which text snippets from a sentence stand for (a) a target word and (b) an instantiation of a frame element.

To identify the target words we simply collected all the words that evoke each frame and implemented a two-step procedure: (1) recognize any of these words in the text sentence; (2) if a word could not be recognized, rank all sentence words by semantic similarity to the evoking words and select the highest ranking word. Semantic similarity is computed with the same procedure employed for generating lexical chains as reported in (Barzilay & Elhadad 1997).

The recognition of FE boundaries is based on a set of heuristics. For example, for the ARRIVING frame, we used a set of 4 heuristics. To describe them, we call *siblings* two phrases that have the same parent in the syntactic parse tree of the sentence being analysed.

- *Heuristic 1* An instantiation of an FE is recognized as an adverbial phrase (ADVP) if:
  (a) The ADVP is a sibling of the target word;
  (b) The head of the ADVP identifies a physical location;
  For example, in the sentence `"Amy arrived home from school early one afternoon."`, Heuristic 1 recognizes [home] as an instantiation of a FE because it is labeled as ADVP by the parser, it is a sibling of the target word *arrive* since they have a common parent (VP) and *home* is a location.
- *Heuristic 2* An instantiation of an FE is recognized as a verb phrase (VP) if:
  (a) The VP is a sibling of the target verb;
  (b) The VP's head is a gerund verb;
  For example, in the sentence `"The Princess of Wales arrived smiling and dancing at a Christmas concert last night."`, Heuristic 2 recognizes the verb phrase `"smiling and dancing"` as a FE instantiation because its head is a gerund verb and a sibling of the target word *arrived*.
- *Heuristic 3* An instantiation of an FE is recognized as a prepositional phrase (PP) in one of the following cases:
  Case 1: (a) PP is a sibling of the target word and (b) the preposition of the PP is *from*, *to via*, *through* or *by*;
  Case 2: (a) PP is a sibling of the target word and (b) the preposition of the PP is *in*, *at* or *on*
  In the previous example, Case 2 of Heuristic 3 recognizes the prepositional phrase `"at a Christmas concert last night"` because it is a sibling of the target word and its preposition is *at*.
- *Heuristic 4* An instantiation of an FE is recognized as a noun phrase (NP) or a wh-phrase (WHNP) [2] if:
  (a) The right-end of the NP or wh-phrase precedes the target word and;
  (b) The NP or wh-phrase are siblings of an ancestor of the target word in the parse tree;
  (c) The NP or the wh-phrase is connected to the target word in the parse tree only through S, SBAR, VP or NP nodes. The NP nodes are allowed only if the target word is of a gerund.
  (d) The NP or the wh-phrase is the top-most and right-most phrase of these types that satisfy conditions (a), (b) and (c).

  For example, in the sentence `"The first of the former concentration camp prisoners and their families will start arriving from the war-torn former Yugoslav republic within days"`, Heuristic 4 recognizes the noun phrase `"The first of the former concentration camp prisoners and their families"` as an instantiation of a FE.

Experiments

The quality of the extraction rules required for any new domain depends on the accuracy with which sentences are labeled with semantic frames relevant to the domain. In our experiments, we generally measured the performance of sentence labeling of a classifier C_a with:

(a) the *precision of labeling*, defined as the ratio between the number of correctly labeled sentences (by C_a) for a frame F_a over the number of sentences processed;

(b) *the recall of labeling* defined as the ratio between the number of sentences correctly labeled with a frame F_a (by C_a) over the number of sentences processed that were labeled (by annotators) for F_a.

(c) The *combined f-measure* defined as $f_1 = \frac{2 \cdot Prec. \cdot Rec.}{Prec. + Rec.}$.

In our tests we have used 9687 sentences from FrameNet annotations, for which the frame labels and all the FE annotations were hidden. Table shows the result of our first pass of the sentence semantic labeling.

[2] a wh-phrase contains a relative pronoun like *who*, *what* or *which*

The table shows the performance of SVM classifiers for 10 frames had the largest number of examples annotated in FrameNet. Precision ranges between 73% and 90%, depending on the semantic frame, whereas recall ranges from 55% to 89%.

In addition; to measure the average performance of the classifiers, we have computed:

1. *the microaverage precision* as the ratio between all sentences labeled correctly for any frame and the total number of processed sentences;
2. *the microaverage recall* as the number of the sentences labeled correctly for any frame over all number of processed sentences;
3. *the microaverage* f_1, computed similarly as the f-measure.

The results listed in Table show that the microaverage f_1 of 80.94% distributed for the entire experiment involving 10 frames. It is close to the f_1 for some of the best-classified frames that lend the largest number of annotations in FrameNet, i.e. JUDGEMENT, MENTAL PROPERTY OR PERCEPTION-NOISE

Table 3: Performance of SVM classifier on frame assignment

Name	Recall	Precision	f_1
self-motion	89.74	87.81	88.76
statement	77.67	80.26	78.94
judgment	83.16	87.36	85.21
perception_noise	75.62	87.18	80.99
experiencer-obj	60.93	80.59	69.39
body-movement	68.56	81.95	74.66
communication_noise	68.74	73.90	71.23
placing	58.06	76.99	66.20
mental-property	79.72	90.81	84.90
leadership	55.89	79.74	65.72
Micro-Average	77.71	84.46	80.94

In each sentence labeled for a frame F_a, we also identify (a) the target word and (b) the boundaries of the FEs that account for the semantic information pertaining F_a. For this purpose we have employed 46 heuristics, many of them applicable across frames that share the same FE. In our experiments, the precision of identification of FEs was 92% while the recall was 78%. When 5624 sentences were processed for the following frames: SELF-MOTION, ARRIVING, DEPARTING and TRANSPORTATION, that we called MV-Frames. From the sentences annotated for MV-Frames, we have identified 285 verbs, called NV-verbs, out of which 158 were target words whereas 127 are verbs identified in the boundaries of FEs. We have identified in the parse trees of the sentences labeled by MV-Frames 285 Subject-Verb-Object triplets.

When applying these new extraction rules to the text evaluated in Event-99, they identified relevant text snippets with a precision of 82% and recall of 58%, thus an F-score or 68%. This result is important because, as reported in (Yangarber *et al.* 2000), if extraction rules perform with high precision, more rules can be learned, thus enhancing the recall. Additionally, the high precision of detecting boundaries of FEs is an essential pre-requisite of semantic parsing of texts, as reported in (Gildea & Jurasky 2002). To our knowledge, this identification is performed manually in current semantic parsers.

Conclusions

We have presented in this paper a new method of automatically acquiring extraction rules for any new domain that can be mapped into a set of semantic frames encoded in FrameNet. In our experiments, the rules obtained performed extraction with high precision, thus enabling full coverage of any new extraction domain when they are further bootstrapped with additional relevant textual information.

This two-pass semantic labeling technique we have developed performs with both human-like precision and recall for a large number of semantic frames. In our experiments we have employed the first release of FrameNet.

References

Barzilay, R., and Elhadad, M. 1997. Using lexical chains for text summarization. In *In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, 1997*.

Fillmore, C. J. 1982. Frame semantics. In *Linguistics in the Morning Calm*, 111–137.

Gildea, D., and Jurasky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistic* 28(3):496–530.

Harabagiu, S., and Maiorano, S. 2000. Acquisition of linguistic patterns for knowledge-based information extraction. In *in Proceedings of LREC-2000, June 2000, Athens Greece*.

Hirschman, L.; Robinson, P.; Ferro, L.; Chinchor, N.; Brown, E.; Grishman, R.; and Sundheim, B. 1999. *Hub-4 Event99 General Guidelines and Templettes*. Springer.

Johnson, C. R., and Fillmore, C. J. 2000. The framenet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *In the Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000), April 29-May 4, 2000, Seattle WA*, 56–62.

McCallum, A. K. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

Riloff, E., and Jones, R. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, 474–479.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer.

Yangarber, R.; Grishman, R.; Tapanainen, P.; and Huttunen, S. 2000. Unsupervised discovery of scenario-level patterns for information extraction. In *Proceedings of the Sixth Conference on Applied Natural Language Processing, (ANLP-NAACL 2000)*, 282–289.