Distributed Neural Embedding for Event Nugget Extraction

Wen Juan Hou
Department of Computer Science
and Information Engineering,
National Taiwan Normal University,
No.88, Sec. 4, Tingzhou Rd.,
Taipei City 116, Taiwan, R.O.C.
emilyhou@csie.ntnu.edu.tw

Bamfa Ceesay
Department of Computer Science
and Information Engineering,
National Taiwan Normal University,
No.88, Sec. 4, Tingzhou Rd.,
Taipei City 116, Taiwan, R.O.C.
bmfceesay@csie.ntnu.edu.tw

Abstract

Given a sentence or an expression, to automatically extract its relevant data and information has gained interest in natural language processing domain. However, it poses a lot of challenging research problems. Extracting events and classifying them into the event types and subtypes gives an additional challenge. This is the TAC 2015 Event Detection challenge. We propose a single model for determination of events, event types, subtypes and REALIS using distributional semantics and the neural embedding techniques.

1 Introduction

There is growing interest in automatic understanding of events and relations between events, or between events and their arguments in the literature. Previous studies categorized the adaptation of techniques for extracting such information from texts into the rule-based and machine learning techniques and the hybrid approach of the two (Chiticariu *et al.*, 2013). It is also argued that the use of the rule-based and machine learning techniques varies in industries and academia. The rule-based information extraction method has enjoyed wide adoption throughout industries due to its explainable ability and the rapid development. However, due to lack of the state-of-the-art approach to formulating rules, the machine learning technique gain much wider adaptation in academia primarily due to its challenging nature (Chiticariu *et al.*, 2010)

In this paper, we use the machine learning technique for event extraction and classification from the literature. Event extraction from texts poses a lot of challenging research problems due to lack of the clear-cut definitions that what an event from a text is. An event can be an explicit occurrence involving participants or a change of a state in place and time. (Mitamura *et al.*, 2015). For efficient extraction of events and classifying events to types and subtypes, our study focuses on semantically meaningful units expressing an event in a sentence. These semantically meaningful units or event nuggets are composed of a single word or a multi-word phrase. We use the definition for types and subtypes by Mitamura *et al.* (2015) and Aguilar *et al.* (2014). As an illustration, consider the following sentences:

- *Several militants were* ***shot dead*** *during clashes near Kabul.* **(a)**

- *The* ***negotiation*** *between the government and the militants was a success.* **(b)**

From the above expressions, the bold face words are event nuggets. It is obvious that, to extract multi-word event nuggets, for example ***shot dead*** as in **(a),** proposes a different challenge from extracting a single word event nugget, for example ***negotiation*** as in **(b)**. We will try to solve this problem in this study.

This paper presents an event extracting technique with the following contributions:

- Extract semantic features using semantic role labeling.

- Predict events, event types and subtypes using Neural Network.

Figure 1 below illustrates a schematic representation of our approach.

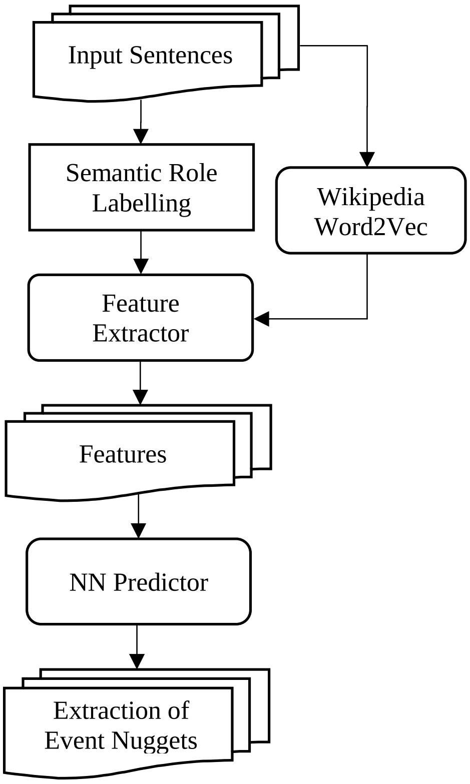

Figure 1: Schematic representation of event nugget extraction

From Figure 1, given the input data, a sentence is preprocessed for semantic role labeling that is an essential part of feature extraction as discussed in Section 2. Wikipedia word2vec is used for semantic embedding. The NN Predictor is the system's skip-gram model for event nugget extraction discussed in section 4.

2 Semantic Role Labeling and Feature Extraction

Semantic Role Labeling (SRL) task is recognizing and labeling semantic arguments of a predicate. A typical semantics includes *Agent*, *Patient*, *Theme*, etc. and it can also include adjunctive arguments indicating *time*, *location*, *manner*, etc. Many semantic representations such as FrameNet (Baker *et al.*, 1998), VerbNet (Schuler, 2006), PropBank (Palmer *et al.*, 2005), etc. are popular for supervised machine learning approaches. This is due to the fact that these corpora are rich with human annotations. There is an extensive use of these corpora in different techniques of semantic role labeling.

Many recent studies show that the information in the syntactic structure of terms in literature can be exploited further for more meaningful feature analysis. A common syntactic information used to extract syntactic features from literatures is the syntactic parse tree. The recent work by Xue and Palmer (2004) proposes feature extraction for semantic role labeling and gives an overview of PropBank corpora and semantic role labeling.

Other studies in the semantic role tagging or identification consider the domain-specific semantic roles such as SPEAKE, MESSAGE, and TOPIC or the abstract semantic roles such as AGENT or PATIENT. A study by Gildea and Jurafsky (2002) uses the statistical classifiers trained on sentences from the FrameNet semantic corpus for further extraction of lexical and syntactic features. The semantic role labeling method has also been used in event extraction problems. A semantic role labeling approach to extracting events from Wikipedia by Exner and Nugues (2011) uses semantic roles (SR) for the event argument identification and property extraction and uses external resources for disambiguation and linking before mapping a predicate structure to an event model. The research on domain-independent detection, extraction and labeling of atomic events (Hatzivassiloglou and Filatova, 2003) has some gains in developing domain independent event extraction system for texts from its atomic level (i.e., sentences and predicates).

In this study we define the semantic role for a given sentence using the probabilistic distribution of features across its syntactic parse structure and use the external corpora to determine disambiguation between roles. A parse structure of a sentence provides rich syntactic relations between lexical terms. However, it can be further processed for the semantic relations (Gildea and Jurafsky, 2002). The following features are used to assign the semantic scores:

Phrase Type. It includes noun phrase (NP), verb phrase (VP) and clause from the parse tree when parsing a sentence. These phrases can be used to express the semantic roles of lexicons.

Grammatical Function. This feature focuses on the parse constituent relations to the rest in a sentence. Only the subject and object relations are considered and we apply it only to NPs because NPs have more effect on the subject and object relations (Gildea and Jurafsky, 2002). NPs with an S (sentence) ancestor are assigned subject roles and NPs with VP ancestors are assigned object roles.

Position. This feature considers if a constituent is before or after a predicate when defining a frame. Generally subjects occur before verbs and objects occur after verbs in an active voice.

Voice. The voice feature refers to the active or passive nature of predicates to capture the connection between the semantic role and the grammatical function.

FrameNet and VerbNet Features. The main idea of FrameNet is that there is a variation to semantic role types available in a particular event. Hence we can constraint the identification of important frames relevant to a particular sentence or predicate to the role searching problem. A generative model for semantic role labeling proposed by Thompson *et al.* (2003) uses the FrameNet corpus for the semantic role and frame identification.

In our study, Framenet and VerbNet features are used to train our model. The target of our system is to identify the event nugget in a given sentence. Therefore given a constituent from a sentence, we have to decide what the semantic type is in respect to FrameNet. This can be determined since the Frame Elements and their associated Lexical Units in FrameNet both reside in the semantic space via frame-to-frame relations and semantic types.

Similarly, VerbNet also provides network structures, revealing relationship such as the sense of application. VerbNet only focuses on lexical terms that are verbs, thus limiting its overall contribution to our model. However, the limitation does not overrun its significance since verb lexicons are sufficiently important in determining the event nugget.

For Training the model, the study uses probabilities calculated for features mentioned above and the details are described in Section 2.1.

2.1 Probability Distribution of Features

This study uses the training data set given in TAC 2015 event nugget track[1]. For training, statistical probabilities are determined across the training data for features mentioned in Section 2 above to train the model in Section 5. As an illustration, given a lexical term *l* and a phrase type *pt*, we can determine the distribution for the semantic role *sr*, as Equation (1):

$$P(sr|l,pt) = \frac{\#(sr,l,pt)}{\#(l,pt)} \tag{1}$$

That is, the probability is calculated as the ratio of the count of each role, #(*sr*,*l*,*pt*), to total number of observation for each conditioning event nugget, #(*l*,*pt*). It is worth to mention that this method significantly works for the simple event nugget with only one lexical term. For the complex event with two lexical terms, we find the joint probability for the two terms of forming the event nugget. An illustration of the probability distribution for the term *"Several"* in the expression *"Several militants were* ***shot dead*** *during clashes near Kabul."* is shown in the table below.

Table 1: An example of probability calculated from the training data

***P*(*sr* | *pt, l, gf*)**	**Scores**
P(*sr=AGT* | *pt=NP, l=Several, gf= Subj*)	0.145
P(*sr=THM* | *pt=NP, l=Several, gf= Subj*)	0.131
P(*sr=THM* | *pt=NP, l =Several, gf=obj*)	0.120
P(*sr=AGT* | *pt=JJ, l=Several*)	0.547
P(*sr=THM* | *pt=JJ, l=Several*)	0.348

In Table 1, *sr* means the semantic role such as agent, AGT, theme, THM, as mentioned in Section 2; *pt* means a phrase type such as NP (noun phrase), JJ (adjective), VB (verb) etc; *l* is a lexicon; and *gf* is the grammatical function defined only for NPs. The *gf* is considered only as subject, *Subj*, or object, *obj* of a sentence.

A similar distribution is defined for FrameNet semantic types and relations that across FrameNet-

[1] http://www.nist.gov/tac/2015/KBP/data.html

VerbNet Mapping[2] from SemLink project[3]. These probabilities are estimated as follows:

$$P(st\ |l, lu, fr) = p(st\ |l, lu) + p(st\ |l, fr) \qquad (2)$$

where

$$P(st\ |l, lu) = \frac{\#(l \in lu)}{\#(lu)} \qquad (3)$$

and

$$p(st|l,, fr) = \frac{\#(l \in fr)}{\#fr} \qquad (4)$$

From the equations above, l is a lexical term, *lu* are the lexical units in a frame and *fr* are frame relations. $\#(l \in lu\ )$ is the number of count *l* in *lu*, $\#(l \in fr)$ is the count of *l* with frame relation in *fr*.

3 Event Nugget Extraction Model

After extensive extraction of features as discussed in Section 2 above, the study utilizes these features to present lexical terms with values through the neural embedding. Consider the following expression:

*Several militants were **shot dead** near Kabul.*

Using the feature values that were extracted, the lexical term “Several”, for example, can be represented as a vector *V*.

$$\boldsymbol{V} = [0.145, 0.131, 0.12, 0.547, 0.348, 0.478] \qquad (5)$$

The first five elements of *V* are *P(r | pt, l, gf)* and the last element is $P(st\ |l, lu, fr)$ value.

Vectors such as *V* are computed for each lexical term and bigram in the input sentence. The study uses the bigram to expand the coverage of the complex event nuggets that are combination of two lexical terms.

Each sentence is represented by a pair set of vectors for single lexical terms and for the bigram of lexical terms respectively. Using the feature vectors for sentences in training data, a skip-gram model is used to learn the distributed vectors.

4 Skip-Gram Model

The training objective of the skip-gram model is to find word representations that are useful for predicting the surrounding words in a sentence (Mikolov *et al*., 2013). Unlike most of other neural network architectures for learning word vectors, the training of the skip-gram model does not involve dense matrix multiplications. Recently an extension of the original skip-gram model propose by Mikolov *et al.* (2013) has shown speedup and accuracy in training and prediction. The diagram below illustrates the skip-gram architecture in our system.

In our model, for a given set of words in a sentence in the training data, the context for a word is in relation to the annotated event nuggets. Figure 2 shows the skip-gram model architecture in our system.

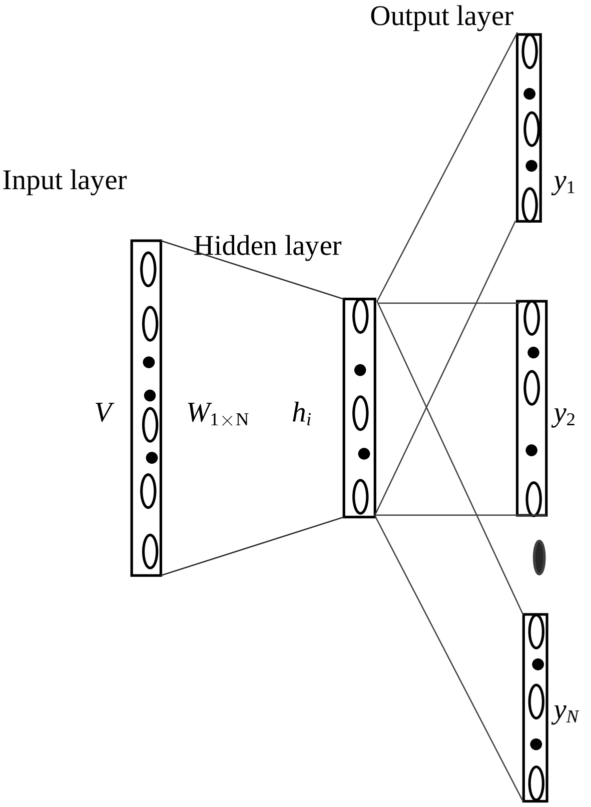

Figure 2: Skip-gram model architecture for learning and predicting event nuggets

In Figure 2, the input of the model is a feature vector of a single lexical term or a bigram *V* and the output is the event nuggets in context of a lexical term ,{y_1, y_2, y_N}, defined in the window size *N*. The hidden layer h_i is composed of vector *W* of 1×N dimension.

[2] https://verbs.colorado.edu/semlink/semlink1.1/vn-fn/

[3] https://verbs.colorado.edu/semlink/

To parameterize the model, the study follows the neural-network language model literature, and models the conditional probability using soft-max as follows:

$$p(c|w;\,\theta) = \frac{e^{Vc.wa}}{\sum_{c'\in C} e^{vc'.wa}} \qquad (6)$$

where vc and wa in R^d are vector representations for c and w respectively. Set C is the set of all available contexts; c is a member of C; and w is a feature vector. The parameter θ is v_c and v_w which are feature and context vectors respectively.

For final determination of event nuggets from this model, only those terms that are assigned output values in relations to the training model are considered as the event nuggets. Therefore, the model performance will be negatively affected by unseen training samples. In handling this problem, Wikipedia corpus is used for the larger scale training in reference to the annotated training data. However, this strategy does not make significant difference, and thus recall is affected.

5 Evaluation and Results

The evaluation system is the tool provided for TAC 2015 Event track [4] evaluation. The evaluation metrics are recall, precision, and F1 score both at micro and macro averages. The evaluation considers four attributes: Event Mention, Types, REALIS Status, and Types and REALIS Status combined as shown in Table 2.

Table 2: Final Mention Detection Results for Event Nuggets Extraction

Macro Average			
Attributes	Precision	Recall	F1-Score
Plain	98.51	63.70	77.37
Type	98.51	63.70	77.37
Realis	98.51	63.70	77.37
Type +Realis	98.51	63.70	77.37
Micro Average			
Attributes	Precision	Recall	F1-Score
Plain	100	55.16	71.10
Type	100	55.16	71.10
Realis	100	55.16	71.10
Type +Realis	100	55.16	71.10

From Table 2, for both macro and micro averages, our system has compromised efficiency for recall. This is due to the weak handling of unseen data during testing.

6 Conclusion

In this study, we present the effectiveness of feature extraction via the semantic role labeling and from the reference corpora. The study also demonstrates a probability distribution across features to translate lexical terms into feature vectors. Neural embedding techniques such as skip-gram in distributional semantics has an extensive usage in speech recognition. While the system reaches good precision, there is still a need to improve its recall. There is a big margin of difference between the recall and the precision.

In the future, there is a plan to continue to improve this work by exploiting more lexical resources for the model's weak point.

References

Aguilar, J., Beller, C., McNamee, P. and Van Durme, B. (2014). A Comparison of the Events and Relations Across ACE, ERE, TAC-KPB, and FrameNet Annotation Standards. *Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation at ACL 2014*, (pp. 45-51).

Baker, C.F., Fillmore, C.J. and Lowe, J.B. (1998). The Berkeley FrameNet Project. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics - Volume 1* (pp. 86-90). Association for Computational Linguistics. doi:10.3115/980845.980860

Chiticariu, L., Li, Y. and Reiss, F.R. (2013). Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 827-832). Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D13-1079

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F. and Vaithyanathan, S. (2010). Domain Adaptation of Rule-based Annotators for Named-entity Recognition Tasks. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1002-1012).

[4] http://www.nist.gov/tac/2015/KBP/Event/index.html

Cambridge, Massachusetts: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1870658.1870756

Exner, P. and Nugues, P. (2011). Using Semantic Role Labeling to Extract Events from Wikipedia. *Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011). Workshop in conjunction with the 10th International Semantic Web Conference*, (pp. 23-24).

FrameNet-VerbNet Mapping. Retrieved from https://verbs.colorado.edu/semlink/semlink1.1/vn-fn/

Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3), 245-288.

Hatzivassiloglou, V. and Filatova, E. (2003). Domain-independent Detection, Extraction, and Labeling of Atomic Events. Columbia University Academic Commons, Retrieved from http://hdl.handle.net/10022/AC:P:20355.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, (pp. 3111-3119).

Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S. and Strassel, S. (2015). Event Nugget Annotation: Processes and Issues. *Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation at the NAACL-HLT 2015.* (pp.66-76).

Palmer, M., Gildea, D. and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), 71-106. Retrieved from http://dx.doi.org/10.1162/0891201053630264

Schuler, K.K. (2006). VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Dissertation. University of Pennsylvania. Retrieved from http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf

SemLink Project. Retrieved from https://verbs.colorado.edu/semlink/

TAC 2015 Event Nugget Track Data. Retrieved from http://www.nist.gov/tac/2015/KBP/data.html

TAC 2015 Event Track. Retrieved from http://www.nist.gov/tac/2015/KBP/Event/index.html

Thompson, C.A., Levy, R. and Manning, C.D. (2003). A Generative Model for Semantic Role Labeling. *Machine Learning: ECML 2003.* (pp. 397-408). Springer Berlin Heidelberg. doi:10.1007/978-3-540-39857-8_36

Xue, N. and Palmer, M. (2004). Calibrating Features for Semantic Role Labeling. *Proceedings of EMNLP 2004*, (pp. 88-94).