Learning to fill the slots from multiple perspectives

Patrik Zajec
patrik.zajec@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Dunja Mladenič
dunja.mladenic@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

We present an approach to train the slot-filling system in a fully automatic, semi-supervised setting on a limited domain of events from Wikipedia using the summaries in different languages. We use the multiple languages and the different topics of the events to provide several alternative views on the data. Our experiments show how such an approach can be used to train the multilingual slot-filling system and increase the performance of a monolingual system.

KEYWORDS

information extraction, slot filling, machine learning, probabilistic soft logic

1 INTRODUCTION

This paper is addressing the slot filling task that aims to extract the structured knowledge from a given set of documents using a model trained for a specific domain and the associated slots. For example, within a news article reporting on an earthquake, the task is to detect the earthquake's magnitude, the number of people injured, the location of the epicentre and other information. We refer to those as a set of *slot keys* or *slots*, to their exact values as a *slot values* and to the named entities from the documents corresponding to those values as *target entities*.

Slot filling is closely related to the task of relation extraction [1] and can be seen as a kind of unary relation extraction. Both tasks can be formulated as classification and are usually approached by first training a classifier with a sentence and tagged entities at the input and the prediction of relation or slot key as the output.

As there is a large number of relations between entities that we might be interested in detecting, there is also a large number of slot keys we seek the slot value for. In order to avoid the resource-intensive process of annotating a large number of examples for each possible slot/relation and to increase the flexibility of training procedures beyond the straight-forward supervised learning, many alternative approaches have been proposed, such as bootstrapping [4], distant supervision [6] and self supervision [5].

As stated both tasks can be performed for different types of documents. We limit our focus to news events on multiple topics (such as natural disasters and terrorist attacks), taking the articles reporting about events as the documents. Since the number of news topics is large, and consequently so is the number of slots, we would like to minimize the need for manual annotations. Furthermore, since the set of topics is not fixed and could expand over time, such a slot filling system should be able to adapt quickly to fill new slots and ideally should not be limited to the English language.

We believe that annotation work can be greatly minimized if we rely on our limited domain to identify and annotate only informative examples and use the additional assumptions to propagate these labels. We also believe that simultaneous training of the system on multiple topics can be advantageous, as we can introduce additional supervision on the common slots and use distinct slots as a source of negative examples.

In this work we use Wikipedia and Wikidata [9] as the source of data. We treat the Wikidata entities that have the point-in-time property specified as events and summary sections of Wikipedia articles about the entity in different languages as news articles. Each entity belongs to a single topic and we adopt the subset of topic-specific properties as slot keys. An automatic exact matching of such values from Wikidata with named entities from Wikipedia articles is rarely successful. We use the successful and unambiguous matches as a set of labeled seed examples.

We formulate the task as a semi-supervised learning problem [8] where the set of base learners is trained iteratively, starting with a small seed set of labeled examples and a larger set of unlabeled examples. In each iteration, the most confident predictions on the examples from unlabeled set are used to increase the training set by assigning pseudo-labels. We introduce an additional component which combines the confidences of multiple base learners for each example.

To the best of our knowledge, we are the first to use the limited domain of news events, which allows the additional assumptions, such as the connection between slots of different topics and the redundancy of reporting in multiple languages, to first train and later boost the performance of a slot-filling system.

The contributions of this paper are the following:

- we combine the data from Wikidata and Wikipedia to setup a learning and evaluation scenario that mimics the learning on news events and articles,
- we demonstrate how simultaneous learning on multiple topics and languages can be used not only to train the multilingual slot-filling system, but to also improve the performance of a monolingual system,
- we show how an inference component can be used to combine predictions from multiple base learners to improve the pseudo-labeling step of the semi-supervised learning process.

2 METHODOLOGY

2.1 Problem Definition

Given a collection of topics $\mathcal{T}$ (such as earthquakes, terrorist attacks, etc.), where each topic t has its own set of slot keys $\mathcal{S}_t$, the goal is to automatically extract values from the relevant texts

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

to fill in the slots. For example, the members of $\mathcal{S}_{earthquakes}$ are *number of injured*, *magnitude* and *location*. For each topic t there is a set of events $\mathcal{E}_t$, each of which took place at some point in time and was reported by several documents in different languages.

The values of all or at least most slot keys (or slots) from $\mathcal{S}_d$ are represented in each of the documents as named entities, which we also refer to as *target entities*. We say most of the slots, since it is possible that an earthquake caused no casualties. It is also possible that some of the documents do not report about the number of casualties as it may be too early to know if there were any. In addition, the documents might contain different values for the same slot key, as for example, the reported number of people injured by an earthquake can increase over time. There may also be several different mentions of the same slot in a particular document, as for example one magnitude might refer to an actual earthquake that the event is about, while the other magnitude might refer to an earthquake that struck the same region years ago.

Our task is actually a two step process. In the first step, the goal is to train a system capable of identifying the target entities for a set of slot keys from the context, which in our case is limited to a single sentence. Such a system is not yet able to recognise the true value for a given slot if there are multiple different candidates, such as selecting the actual magnitude from several reported magnitude values. The goal of the second step is to assign a single correct value to each of the slot keys. We assume that inferring the correctness of a value is a document-level task, since it requires a broader context. Solving the first step is a kind of prerequisite for the second step, so we focus on it in this paper.

2.2 Overview of the proposed method

The system is trained iteratively and starts with a noisy seed set, which grows larger with pseudo-labeled positive and negative examples. Each of the base learners is trained on the set of labeled examples from the topic (or multiple topics) and language assigned to it. The prediction probabilities for each of the unlabeled examples are determined by combining the probabilities of all base learners. This is done either by averaging or by feeding the probabilities as approximations of the true labels into the component, which attempts to derive the true value for each example and the error rates for each learner [7]. The examples with probabilities above or below the specific thresholds are given a pseudo-label and added to the training set.

The seed set is constructed by matching the slot values obtained from Wikidata with named entities found in Wikipedia articles for each event. There are only a handful of unambiguous matches for each slot key, which are labeled as a positive examples, while the negative examples are all other named entities from the articles in which they appeared. Figure 1 shows a high-level overview of the proposed methodology. The entire workflow is repeated in each iteration until no new examples are selected for pseudo-labelling.

2.3 Representing the entities

Each named entity together with its context forms a single example. We annotate each article and extract the named entities with Spacy [1]. To capture the context, we compute the vector representation of each entity by replacing it with a mask token and feeding the entire sentence through a pre-trained version of the XLM Roberta model [3] using the implementation from the Transformers [2] library. Note that the representation of each entity remains fixed throughout the learning process because we have found that the representation is expressive enough for our purposes and it speeds up the training between iterations. Also note that since the entity is masked, it is not directly captured in the representation.

2.4 Selecting the topics

Our assumption is that training the system to detect the slots on multiple topics simultaneously can provide additional benefits. For two topics t and t' there is potentially a set of common slots and a set of topic-specific slots.

For slot s' which appears in both topics the base learner trained on t' can be used to make predictions for examples from t. By combining predictions from learners trained on t and t', we could get a better estimate of the true labels of the examples.

For the slot s, which is specific to the topic t, all examples from the topic t' can be used as negative examples. Selecting reliable negative examples from the same topic is not easy, as we may inadvertently mislabel some of the positive examples.

2.5 Using multiple languages

Articles from different languages offer in some ways different views on the same event. The slot values we are trying to detect should appear in all the articles, as they are highly relevant to the event.

The values for slots such as location and time should be consistent across all articles, whereas this does not necessarily apply to other slots such as the number of injured or the number of casualties. Matching such values across the articles is therefore not a trivial task, and although a variant of soft matching can be performed, we leave it for the future work and limit our focus only on the values that can be matched unambiguously.

We can combine the predictions of several language-specific base learners into a single pseudo-label for entities that can be matched across the articles.

2.6 Assigning pseudo labels

Each iteration starts with a set of labeled examples X_l, a set of unlabeled examples X_u and a set of base learners trained on X_l. Base learners are simple logistic regression classifiers that use vector representations of entities as features and classify each example x as a target entity for the slot key s or not.

Each base learner $\bar{f}^s_{t,l}$ is a binary classifier trained on the labeled data for the slot key s from the topic t and the language l. Such base learners are *topic-specific* as they are trained on a single topic t. Base learners $\bar{f}^s_l$ are trained on the labeled data for the slot key s from the language l and all the topics with the slot key s. Such base learners are *shared* across topics, as they consider the examples from all the topics as a single training set. We use the classification probability of the positive class instead of hard labels, $\bar{f}^s_{t,l}(x), \bar{f}^s_l(x) \in [0, 1]$.

For each entity x from a news article with the language l reporting on the event e from the topic t we obtain the following predictions:

- $\bar{f}^s_{t',l}(x)$ for each $s \in \mathcal{S}_t$ and all such t' that $s \in \mathcal{S}_{t'}$, that is the probability that x is a target entity for the slot key

[1] https://spacy.io/

[2] https://huggingface.co/transformers/

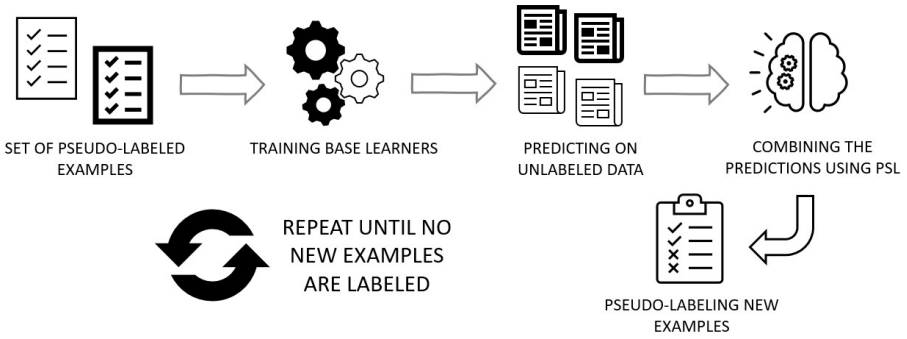

Figure 1: High-level overview of the proposed methodology.

s, where s is a slot key from the topic t, using the *topic-specific* base learner trained on examples from the same language on the topic t' that also has the slot key s,

- $\bar{f}^s_{t,l'}(x)$ which equals $\bar{f}^s_{t,l'}(y)$ for each $s \in \mathcal{S}_t$ and for each language l' such that there is an article reporting about the same event e in that language and contains an entity y which is matched to x,
- $\bar{f}^s_l(x)$ for each $s \in \mathcal{S}_t$, using the *shared* base learner, which is on examples from all topics t' that have the slot key s.

Predictions from multiple base learners for each x and s are combined as a weighted average to obtain a single prediction $f^s(x)$. The weight of each base learner $\bar{f}$ is determined by its error rate $e(\bar{f})$ which is estimated using an approach from [7] using both unlabeled and labeled examples. This is done by introducing the following logical rules (referred to as *ensemble rules* in [7]) for each of the base learners $\bar{f}^s$ predicting for x:

$$\bar{f}^s(x) \wedge \neg e(\bar{f}^s) \rightarrow f^s(x),\ and,\ \bar{f}^s(x) \wedge e(\bar{f}^s) \rightarrow \neg f^s(x),$$
$$\neg\bar{f}^s(x) \wedge \neg e(\bar{f}^s) \rightarrow \neg f^s(x),\ and,\ \neg\bar{f}^s(x) \wedge e(\bar{f}^s) \rightarrow f^s(x).$$

The truth values are not limited to Boolean values, but instead represent the probability that the corresponding ground predicate or rule is true. For a detailed explanation of the method we refer the reader to [7]. We introduce a prior belief that the predictions of base learners are correct via the following two rules:

$$\bar{f}^s(x) \rightarrow f^s(x),\ and,\ \neg\bar{f}^s(x) \rightarrow \neg f^s(x).$$

Since each x can be target entity for at most one slot key, we introduce a *mutual exclusion* rule:

$$\bar{f}^s(x) \wedge f^{s'}(x) \rightarrow e(\bar{f}^s).$$

The rules are written in the syntax of a Probabilistic soft logic [2] program, where each rule is assigned a weight. We assign a weight of 1 to all *ensemble rules*, a weight of 0.1 to all *prior belief* rules and a weight of 1 to all *mutual exclusion* rules. The inference is performed using the PSL framework [3]. As we obtain the approximations for all $x \in X_u$, we extend the set of positive examples for each slot s with all x such that $f^s(x) >= T_p$ and the set of negative examples with all x such that $f^s(x) <= T_n$, for predefined thresholds T_p and T_n.

3 EXPERIMENTS

3.1 Dataset

To evaluate the proposed methodology, we have conducted experiments on two topics: *earthquakes* and *terrorist attacks*.

[3]https://psl.linqs.org/

We have collected the Wikipedia articles and Wikidata information of 913 earthquakes from 2000 to 2020 in 6 different languages, namely English, Spanish, German, French, Italian and Dutch. We have manually annotated the entities of 85 English articles using the slot keys *number of deaths*, (*number of injured* and *magnitude*, which serve as a labeled test set and are not included in the training process. In addition, we have collected the data of 315 terrorist attacks from 2000 to 2020 with the articles from the same 6 languages.

3.2 Evaluation Settings

The evaluation for each approach is performed on the labeled English dataset, where 76 entities are labeled as number of deaths, 45 as number of injured and 125 as magnitude. The threshold values for the pseudo-labeling are set to $T_p = 0.6$ and $T_n = 0.05$. The approaches differ by the subset of base learners used to form the combined prediction and by the weighting of the predictions.

Single or multiple languages. In single language setting, only English articles are used to extract the entities and train the base learners. In the multi-language setting, all available articles are used and the entities are matched across the articles from the same event.

Single or multiple topics. In the single topic setting only the examples from the *earthquake* topic are used. In the multi-topic setting, the examples from *terrorist attacks* are used as negative examples for the slot key *magnitude*, the base learners for the slot keys *number of deaths* and *number of injured* are combined as described in the section 2.6.

Uniform or estimated weights. In the uniform setting all predictions of the base learners contribute equally, while in the estimated setting the weights of the base learners are estimated using the approach described in the section 2.6.

3.3 Results and discussion

The results of all experiments are summarized in the table 1. Since the test set is limited to the topic *earthquake* and English, only a subset of base learners was used to make the final predictions. We report the average value of precision, recall and F1 across all slot keys. The threshold of 0.5 was used to round the classification probabilities.

Single iteration. Approaches in which base learners are trained on the initial seed set for a single iteration achieve higher precision with the cost of a lower recall. We observe that they distinguish almost perfectly between the slots from the seed set and

Table 1: Results of all experiments. The column *Single iteration* reports the results of approaches where base learners were trained on the seed set only. Results where base learners were trained in the semi-supervised setting with different weightings of the predictions are reported in the columns *Uniform weights* and *Estimated weights*. The values of precision, recall and F1 are averaged over all slot keys.

	Single iteration			Uniform weights			Estimated weights		
Model	P	R	F1	P	R	F1	P	R	F1
Single language, single topic	0.94	0.64	0.76	0.83	0.75	0.77	0.84	0.76	0.79
Multiple languages, single topic	0.94	0.64	0.76	0.82	0.74	0.76	0.83	0.75	0.77
Single language, multiple topics	0.91	0.76	0.83	0.83	0.83	0.83	0.86	0.83	0.84
Multiple languages, multiple topics	0.93	0.76	0.83	0.82	0.83	0.82	0.84	0.84	0.84

produce almost no false positives. Using one or more languages has almost no effect on the averaged scores when the number of topics is fixed. When using multiple topics, a higher recall is achieved without a significant decrease in precision. All incorrect classifications of the slot *number on injured* are actually examples of the *number of missing* slot that is not included in our set and likewise almost all incorrect classifications for the slot *magnitude* are examples of the slot *intensity on the Mercalli scale*. This could easily be solved by expanding the set of slot keys and shows how important it is to learn to classify multiple slots simultaneously.

Semi-supervised. Approaches in which base learners are trained iteratively trade precision in order to significantly improve recall. Most of the loss of precision is due to misclassification between slots *number of deaths* and *number of injured*, similar as the example *"370 people were killed by the earthquake and related building collapses, including **228** in Mexico City, and more than 6,000 were injured."* where 228 was incorrectly classified as number of injured and not the number of deaths. The use of multiple topics reduces misclassification between these slots and further improves the recall as new contexts are discovered by the base learners trained on *terrorist attacks*.

Uniform and estimated weights. Using the estimated error rates as weights for the predictions of base learners shows a slight improvement in performance. It may be advantageous to estimate multiple error rates for *topic-specific* base learners, as they tend to be more reliable in predicting examples from the same topic. We believe that more data and experimentation is needed to properly evaluate this component. A major advantage is its flexibility, since we can easily incorporate prior knowledge of the slots or additional constraints on the predictions in the form of logical rules.

4 CONCLUSION AND FUTURE WORK

We presented an approach for training the slot-filling system which can benefit from large amounts of data from Wikipedia. The experiments were performed on a relatively small dataset and show that the proposed direction seems promising. However, the right test of our approach would be to apply it to a much larger number of topics and events, which will be done in the immediate next step. Furthermore, the current approach needs to be evaluated in more detail.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and NAIADES European Unions project under grant agreement H2020-SC5-820985.

REFERENCES

[1] Nguyen Bach and Sameer Badaskar. 2007. A Survey on Relation Extraction. Technical report. Language Technologies Institute, Carnegie Mellon University.

[2] Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss markov random fields and probabilistic soft logic. *The Journal of Machine Learning Research*, 18, 1, 3846–3912.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[4] Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning. In *Proceedings of AAAI*.

[5] Xu ming Hu, Lijie Wen, Y. Xu, Chenwei Zhang, and Philip S. Yu. 2020. Selfore: self-supervised relational feature learning for open relation extraction. *ArXiv*, abs/2004.02438.

[6] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 1003–1011.

[7] Emmanouil Platanios, Hoifung Poon, Tom M Mitchell, and Eric J Horvitz. 2017. Estimating accuracy from unlabeled data: a probabilistic logic approach. In *Advances in Neural Information Processing Systems*, 4361–4370.

[8] Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109, 2, 373–440.

[9] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57, 10, 78–85.