UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Prior-informed distant supervision for temporal evidence classification

Reinanda, R.; de Rijke, M.

Publication date
2014
Document Version
Final published version
Published in
COLING 2014: the 25th International Conference on Computational Linguistics

Link to publication

Citation for published version (APA):
Reinanda, R., & de Rijke, M. (2014). Prior-informed distant supervision for temporal evidence classification. In J. Tsujii, & J. Hajic (Eds.), *COLING 2014: the 25th International Conference on Computational Linguistics: proceedings of COLING 2014 : technical papers: August 23-29, 2014, Dublin, Ireland* (pp. 996-1006). Association for Computational Linguistics. http://www.aclweb.org/anthology/C/C14/C14-1094.pdf

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:08 Oct 2026

Prior-informed Distant Supervision for Temporal Evidence Classification

Ridho Reinanda
University of Amsterdam
Amsterdam, The Netherlands
r.reinanda@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

Abstract

Temporal evidence classification, i.e., finding associations between temporal expressions and relations expressed in text, is an important part of temporal relation extraction. To capture the variations found in this setting, we employ a distant supervision approach, modeling the task as multi-class text classification. There are two main challenges with distant supervision: (1) noise generated by incorrect heuristic labeling, and (2) distribution mismatch between the target and distant supervision examples. We are particularly interested in addressing the second problem and propose a sampling approach to handle the distribution mismatch. Our prior-informed distant supervision approach improves over basic distant supervision and outperforms a purely supervised approach when evaluated on TAC-KBP data, both on classification and end-to-end metrics.

1 Introduction

Temporal relation extraction is the problem of extracting the temporal extent of relations between entities. A typical solution to the temporal relation extraction problem has three main components: (1) *passage retrieval*, (2) *temporal evidence classification*, and (3) *temporal evidence aggregation*. A community-based effort to evaluate temporal relation extraction was introduced in 2011 as a TAC Knowledge Base Population task: Temporal Slot Filling, or TSF for short (Ji et al., 2011).

An illustration of temporal slot filling is as follows. Having identified a `per:spouse` relation between two entities (Freeman Dyson, Imme Dyson), a system must establish the temporal boundaries from its supporting sentence. In the case of the sentence "*In 1958, he married Imme Dyson*", the goal is to find that the relation lasts from 1958 until the present day. Within the TSF setting, the boundaries are represented as beginning and ending intervals in a tuple (T_1, T_2, T_3, T_4) instead of an exact time expression, so as to allow uncertainty in the system output. We investigate temporal relation extraction following this setting. We focus on the temporal evidence classification part.

One of the challenges with relation extraction is the limited amount of training data available to capture the variations in a target corpus: temporal relation extraction faces the same challenge. Employing distant supervision (Mintz et al., 2009) is a way to address the challenge. But generating example training data in the temporal setting is not straightforward: we have to find not only the query and related entity, but also the time expression, in a single text segment.

Employing distant supervision for temporal evidence classification will introduce noise, in the form of labels and additional contexts (e.g., lexical features). A lot of previous work in distant supervision has been dedicated to reducing noise in distant supervision (Bunescu and Mooney, 2007; Riedel et al., 2010; Wei et al., 2012). We are interested in another phenomenon: the class distributions found in training data generated by a distant supervision approach. These distributions become an issue if the distant supervision corpus has a different structure and different characteristics compared to the target corpus, e.g., Wikipedia vs. news articles. We observe that in the case of temporal evidence, news articles and Wikipedia do indeed contain different class distributions. Our working hypothesis is that incorporating prior information about temporal class distribution helps improve our distant supervision approach. We

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: `http://creativecommons.org/licenses/by/4.0/`

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 996–1006, Dublin, Ireland, August 23-29 2014.

test this hypothesis by comparing a distant supervision strategy with class priors to a distant supervision without class priors. We also demonstrate the effectiveness of our method by contrasting it with a purely supervised approach. In addition, we investigate how the difference in performance in temporal evidence classification affects the final score obtained in the overall end-to-end task.

We discuss related work in Section 2. In Section 3, we describe our distant supervision approach for temporal evidence classification. Our experimental setup is detailed in Section 4. We follow with results in Section 5 and a conclusion in Section 6.

2 Related Work

We discuss two groups of related work: on temporal slot filling and on distant supervision.

2.1 Temporal slot filling

Some previous work uses a pattern-based approach (Byrne and Dunnion, 2011); patterns are defined in terms of query entity, temporal expression, and slot value. For example, the word *divorce* should trigger that the relation *per:spouse* is ending. Other work uses temporal linking between time expressions and events in an event-based approach (Burman et al., 2011), where the source documents are annotated with TimeML event annotations (Pustejovsky et al., 2003); the authors use intra-sentence event-time links, and inter-sentence event-event links, following a TempEval approach (UzZaman et al., 2012). Garrido et al. (2012) use a graph-based document representation; they convert document context to a graph representation and use TARSQI to determine links between time expressions and events in documents and later map the resulting links into five temporal classes.

Li et al. (2012) combine flat and structured approaches to perform temporal classification. Their approach relies on a custom SVM kernel designed around flat (window and shallow dependency) features and structured (dependency path) features. The structured approach is designed to overcome the long context problem. They use a distant supervision approach for the temporal classification part, obtained on Freebase relations. They further extend their approach with self-training and relabeling (Ji et al., 2013).

Finally, Surdeanu et al. (2011) use n-grams around temporal expressions to train a distant supervision system. To be able to use Freebase facts, they find example sentences in Wikipedia, and use a window of five words from the temporal expression, using Freebase facts as *start* and *end* trigger. They use Jaccard correlation between n-grams to determine the association to *start* and *end*. Sil and Cucerzan (2014) performed distant supervision using facts obtained from Wikipedia infoboxes. From Wikipedia infoboxes, they retrieve the relevant sentences and build n-gram language models of the relations. In a slightly different setting (exploratory search), Reinanda et al. (2013) establish the temporal extent of entity associations simply by looking at their co-occurrence within documents in the corpus.

Our approach to temporal evidence classification differs from most existing approaches in its distant supervision scheme. We use distant supervision to directly perform a multi-class classification of temporal evidence against the five main temporal classes (including the *before* and *after* class), where most of the previous systems train a model to detect the beginning and ending of relationships only.

2.2 Reducing noise in distant supervision

With distant supervision (Mintz et al., 2009), indirect examples in the form of relations from a knowledge base such as Freebase and DBPedia are used. From these relation tuples, instances of relations in the form of sentences in the corpus are searched. Text features are later extracted from these sentences that are then used to train classifiers that can identify relations in the text corpus.

Reducing noise is an important ingredient when working with a distant supervision assumption. Relabeling is one such approach; Tamang and Ji (2012) perform relabeling based on semi-supervised lasso regression to reduce incorrect labeling. Wei et al. (2012) show that instances may be labeled incorrectly due to the knowledge base being incomplete. They propose to overcome the problem of incomplete knowledge bases for distant supervision through passage retrieval model with relation extraction.

Ritter et al. (2003) focus on the issue of missing data for texts that contain rare entities that do not exist in the original knowledge base. Riedel et al. (2010) work with a relaxed distant supervision assumption; they design a factor graph to explicitly model whether two entities are related, and later train this model with a semi-supervised constraint-driven algorithm; they achieve a 31 percent error reduction.

Bunescu and Mooney (2007) introduce multiple instance learning to handle the weak confidence in the assigned label. They divide the instances into a positive bag (at least one positive example) and a negative bag (all negative examples). They design a custom kernel to work with this weaker form of supervision. Surdeanu et al. (2012) operate on the same principle, but model the relation between entities and relation classes using graphical models. Hoffmann et al. (2011) also use multi-instance learning, but focus on overlapping relations.

What we add on top of existing work is the use of sampling techniques to correct for skewed distributions introduced through distant examples. We propose prior sampling, correcting the distributions of the classes in the generated examples to fit the target corpora.

3 Method

The temporal slot filling task is defined as follows: given a relation R = (q, r, s), where q is a query entity, r is a related entity, and s is a slot type, one must find T_R, a tuple of four dates (T_1, T_2, T_3, T_4) where R holds, where T_1 and T_2 form the beginning interval of the relation, and T_3 and T_4 is the ending interval. A system first must retrieve all passages or sentences expressing the relation between q and r. Each sentences and any time information within them will serve as intermediate evidence. This temporal evidence will later be aggregated and converted to tuple representation T_R.

In this paper, we focus on *temporal evidence classification*. That is, assuming the passage retrieval component has retrieved the relevant passages as intermediate evidence of temporal relations, we must classify whether the time expression t in the passage belongs to one these classes: BEGINNING, ENDING, BEFORE, AFTER, and WITHIN. In the training and evaluation data available to us, only the offsets of the time expression within the document are given for each intermediate evidence, therefore we first extract the paragraph and find the context sentence mentioning t.

Distant supervision for temporal classification The temporal slot filling task, as specified by TAC-KBP, defines 7 types of temporal-intensive relations. In our distant supervision approach, we use a separate knowledge base to find instances of the equivalent relations. We use Freebase as our reference knowledge base. That is, we use the temporal information found in Freebase to generate training examples. We manually map the TAC-KBP's 8 temporal relations into 6 Freebase mediator relations. The complete mapping of the relations can be found in Table 1.

In an article, entities and time expressions are not always referred to using their full mentions within a single sentence. Sometimes information is scattered around several sentences: the query entity q in the first sentence, later referred to using a pronoun in the second sentence that contains a time expression, etc. One common way to deal with this problem is to run full co-reference resolution, therefore ensuring all mentions are resolved. We handle this problem by relaxing the distant supervision rule. Rather than retrieving sentences, we retrieve passages containing the query entity q, and related entity r instead. We later replace every pronoun found within the passage with q. Based on our analysis of the Wikipedia articles, this simple heuristic should work, because most Wikipedia articles are entity-centric, and a lot of the pronouns mentioned in the articles will refer to the query entity q.

Each relation that we mapped from Freebase has temporal boundaries *from* and *to*. Following Li et al. (2012), we use Algorithm 1 to generate the training examples, but adapt it suit to our assumption.

Sampling the DS examples We manually compared our main corpus (TAC document collection) and our distant supervision corpus (Wikipedia) and noticed some discrepancies. The main corpus mainly consists of newswire articles; one of the main difference between Wikipedia articles and newswire articles is that Wikipedia articles mainly consist of milestone events. In terms of class distribution, this means that most of the generated examples will be in the form of BEGINNING and ENDING class, followed by the BEFORE and AFTER class, with the smallest number of examples belonging to the

TAC Relations	Freebase Relations
per:spouse	marriage
per:title	employment-tenure, goverment-position-held
per:employee-of	employment-tenure
per:member-of	political-party-tenure
per:cities-of-residence	places-lived
per:stateorprovinces-of-residence	places-lived
per:countries-of-residence	places-lived
org:top-employees/members	organization-leadership

Table 1: Relation mapping to Freebase.

Data: Freebase temporal relation (q, r, $from$, to)
Result: labeled training examples
Retrieve the Wikipedia article of the query entity q;
Split article into passages;
Retrieve the passages containing q, r;
Extract all time expressions from the passages;
for *time-expression* t **do**
    Retrieve the context sentence s containing t;
    If t is $from$: use s, t as BEGINNING example;
    If t is to : use s, t as ENDING example;
    If t before $from$: use s, t as BEFORE example;
    If t after to : use s, t as AFTER example;
    If t between $from$ and to : use s, t as WITHIN example;
end

Algorithm 1: Training data generation.

WITHIN class. In newswire, however, we tend to see something different; most of the time expressions will belong to the WITHIN class.

We argue that using the training data with a "smarter" prior is important. More data not only means more information, but may also mean more noise. This is particularly important with the *relaxed distant supervision* assumption that we have. Therefore, we choose to sample instead of using all of the generated training examples.

We employ two sampling strategies: *uniform*, sampling from our generated training data and deliberately fitting them to a uniform distribution; and *prior-sampling*, where we deliberately construct training data to fit a prior distribution. One way to estimate such a prior is by looking at the distributions of classes in the gold-standard training data that we have. In the case where gold-standard data is not available, we can use a heuristic to estimate the distributions of temporal classes based on domain knowledge or on observations of the target corpora.

In summary, we generate the final training data according to the following steps. First, generate training data with the DS approach described before. Next, estimate class distributions from the (supervised) training data. Then, sample examples from the generated DS data with the probability estimated from the supervised training data (i.e., the empirical prior). Keep sampling the training examples until we

reach the target percentage of the DS data. Finally, use the sampled training data to train the multi-class classifier.

Feature representation Both for the training, evaluation, and DS data, we extract the context sentence, i..e, the sentence containing the relation and time expression t.

We normalize the context sentence as follows. First, we detect named entities within the sentence and replace the mentions with their entity types (PERSON, ORGANIZATION, or LOCATION). Second, we detect other time expressions within the context and normalize them with regard to the main time expression t, i.e., by normalizing them into TIME-LT and TIME-GT. The idea is to capture the relationships between time expressions as features.

We extract lexical features from the normalized sentence. This comprises tokens surrounding the query entity, related entity (slot filler), and time expression. We consider the following four models as our feature representations:

Model-1: bag-of-words All tokens within the normalized sentences are used as features.

Model-2: context window All tokens within the proximity of 3 tokens from the query entity, related entity, and time expression are used as features.

Model-3: context window with trigger words lexicon All tokens within the proximity of 3 token from the query entity, related entity, and time expression are used as features. In addition, a list of keywords which might indicate the beginning and ending of relationships are used as gazetteer features. These list of keywords are expanded by using WordNet to extract related terms.

Model-4: context window with position All tokens within the proximity of 3 tokens from the query entity, related entity, and time expression are used as features. Rather than considered as bag-of-words tokens, the positions of word occurrences are now taken into account as features.

4 Experimental Setup

We introduce the dataset and the setup of our experiments. Before that we formulate our research questions as these dictate our further choices.

Research questions We aim to answer the following research questions:

RQ1 How does a purely supervised approach with different features and learning algorithms perform on the task of temporal evidence classification?

RQ2 How does the performance of a distant supervision approach compare to that of a supervised learning approach on the task of temporal evidence classification?

RQ3 How does the performance of a prior-informed distant supervision approach compare to that of a basic distant supervision approach on the task of temporal evidence classification?

RQ4 How do the approaches listed above compare in terms of their performance on the end-to-end temporal relation extraction task?

Corpora and knowledge base We use the TAC 2011 document collection, which contains 1.7M documents, consisting of news wires, web texts, broadcast news, and broadcast conversation. We use a recent version of Freebase (October 2013) as our knowledge base and retrieve the latest version of Wikipedia as our distant supervision corpus.

Ground truth We use the TAC-KBP 2011 Temporal Slot Filling Task dataset (Ji et al., 2011) as the ground truth in our experiments. The ground truth comes in two forms: intermediate evidence (with classification labels) and tuples (boundaries of each relation). We use the intermediate evidence to evaluate our temporal evidence classification framework. We later use the provided tuples to evaluate the end-to-end result.

The dataset contains 173 examples in the training set and 757 examples in the evaluation set. The distribution of the classes is shown in Table 2.

Class	Training	Evaluation	DS Training
WITHIN	66	357	6,129
BEGINNING	59	217	22,508
ENDING	30	110	16,775
BEFORE	9	45	24,932
AFTER	9	28	12,499

Table 2: Class distribution statistics.

Evaluation metric We use F1 as the main evaluation metric for the temporal evidence classification task. For the end-to-end temporal information extraction task, we use the evaluation metric proposed in TAC-KBP 2011, i.e., the Q score. Given a relation r and the ground truth interval tuple G_r, $Q(T_r)$, the quality score of a tuple T_r returned by system S is computed as follows:

$$Q(T_r) = \frac{1}{4} \sum_{i=1}^{4} \frac{1}{1+d_i},$$

where d_i is the absolute difference between T_i in system response and the ground truth tuple G_i (measured in years). To obtain an overall system Q score, we average the Q scores obtained from each relation tuple returned.

Experiments We run four contrastive experiments. In Experiment 1, we contrast the performance on the temporal evidence classification task of the different choices for our supervised methods (Model-1, -2, -3, -4), using either Support Vector Machine, Naive Bayes, Random Forest, or Gradient Boosted Regression Tree. In Experiment 2 we examine our distant supervision method and contrast its performance with the supervised methods from Experiment 1. In Experiment 3, we contrast different sampling methods for our distant supervision method.

In Experiment 4 we consider the overall performance on the temporal relation extraction task of our methods; in this experiment we use three "oracle runs" that we have not introduced yet: first, the *Label-Oracle* run uses the actual temporal classification label from the ground truth, use these ground truth label to aggregate the evidence and create the temporal tuples, and compute the end-to-end score; second, *Within-Oracle* assigns all temporal evidence to the WITHIN class; third, *Nil-Baseline* is a lower-bound run that assigns NIL to every element of the temporal tuples.

We use the implementations of the learning algorithms in the Scikit-learn machine learning package (Pedregosa et al., 2011).

5 Results and Discussion

We present the outcomes of the four experiments specified in the previous section.

5.1 Preliminary experiment

To answer RQ1, *How does the performance of the supervised learning approaches on the temporal evidence classification task vary with different representations and learning algorithms?*, we start with a preliminary experiment. The aim of this experiment is to get an idea of the classification performance with a purely supervised approach. The results are shown in Table 3.

As shown in Table 3, Model-4 with the SVM and NB classifiers achieves the best overall performance. There seems to be a gradual increase in performance from the simpler to the more complex model with SVM and NB classifiers, with the exception of RF. Interestingly, GBRT seems only slightly affected by the different choice of model in this supervised setting.

5.2 Distant supervision experiments

Next, we evaluate the distant supervision approach. We aim to answer RQ2, *How does the performance of the distant supervision approach compare to that of the supervised learning approach?* We generate

Model	SVM	NB	RF	GBRT
Model-1	0.405	0.361	0.402	0.422
Model-2	0.409	0.417	0.354	0.420
Model-3	0.412	0.418	0.361	0.420
Model-4	**0.426**	0.424	0.241	0.422

Table 3: Experiment 1. Supervised approaches to temporal evidence classification.

training examples with the approach described in Section 3, and use the full generated training data to train SVM and Naive Bayes classifiers with the same representation models that we use in the previous experiments. The results are shown in Table 4.

Model	Supervised	DS	DS-uniform	DS-prior
Model-1 SVM	0.405	0.212	0.379	0.408
Model-2 SVM	0.409	0.185	0.389	0.450
Model-3 SVM	0.412	0.183	0.384	0.452
Model-4 SVM	0.426	0.200	0.400	0.463
Model-1 NB	0.361	0.413	0.379	0.431
Model-2 NB	0.417	0.299	0.372	0.451
Model-3 NB	0.418	0.300	0.368	0.446
Model-4 NB	0.424	0.270	0.400	**0.486**
Model-1 RF	0.402	0.162	0.406	0.397
Model-2 RF	0.354	0.177	0.399	0.418
Model-3 RF	0.361	0.176	0.391	0.403
Model-4 RF	0.241	0.171	0.399	0.446
Model-1 GBRT	0.422	0.142	0.316	0.344
Model-2 GBRT	0.420	0.137	0.343	0.418
Model-3 GBRT	0.420	0.138	0.343	0.403
Model-4 GBRT	0.422	0.140	0.399	0.433

Table 4: Experiment 2 and 3. Supervised, distant supervision, and distant supervision with sampling approaches to temporal evidence classification.

We observe that the distant supervision approach trained on the full set of generated examples (the column labeled "DS") performs poorly, well below the supervised approach. We hypothesize that the accuracy drops due to the amount of noise generated with our distant supervision assumption trained from full data, and different class distribution statistics.

In Section 3, we proposed our prior-sampling approach for distant supervision. The next experiment is meant to answer RQ3, *How does the performance of our prior-informed distant supervision approach compare to that of the basic distant supervision approaches?* We sample 20 percent of the generated examples datasets with the following strategies: *uniform* and *prior*. The results are also shown in Table 4, in the columns labeled "DS-uniform" and "DS-prior," respectively.

By observing the results in Table 4, we notice that distant supervision with prior sampling performs the best, for every combination of model and classification method. *Uniform* sampling already helps in improving the performance, and prior sampling successfully boosts the performance of the basic distant supervision (for all four models) further. Distant supervision with prior sampling also performs consistently better than the supervised approaches (Table 3) in many cases—interestingly, for GBRT, DS-prior only outperforms the supervised methods with sufficiently complex queries (Model-4 GBRT).

5.3 End-to-end experiments

Next, we answer RQ4. That is, we consider how the classification performance on temporal evidence classification affects the end-to-end result. We take the best performing models from the previous experiments and evaluate their end-to-end scores. The results are shown in Table 5.[1]

Model	Avg-Q	F1
Label-Oracle	0.925	1.000
Within-Oracle	0.676	0.302
Nil-Baseline	0.393	N/A
Supervised		
Model-4 SVM	0.657	0.426
Model-4 NB	0.648	0.424
Model-4 RF	0.573	0.241
Model-4 GBRT	0.649	0.422
Distant supervision		
Model-4 SVM	0.669	0.463
Model-4 NB	0.679	0.486
Model-4 RF	0.653	0.446
Model-4 GBRT	0.669	0.433

Table 5: Experiment 4. End-to-end scores (Avg-Q) next to F1 scores for temporal evidence classification.

From Table 5, we see that Model-4 RF (F1 on temporal evidence classification 0.446) and Model-4 GBRT (F1 on temporal evidence classification 0.433) translate into 0.653 and 0.669, respectively, in terms of Q-score. This means that the misclassifications that Model-4 RF produces have a larger impact than those of Model-4 GBRT. However, the difference in performance is not large.

The evaluation of this end-to-end task is important because not every misclassification has a similar cost. Misclassification of class A into class B can result in a huge increase/decrease in performance. First, the classification performance does not directly map to the end-to-end score. Second, several relations have more pieces of evidence than others; performing misclassifications on relations that have a lot of supporting evidence would probably have less effect on the final score.

The state of the art performance, using distant supervision (Li et al., 2012), achieves an end-to-end Avg-Q score of 0.678 (on training data), where we achieve 0.679 (on evaluation data). However, our scores are not directly comparable since we reduce the number of classes (and the amount of evidence) in our evaluation. It is important to note that Li et al. (2012) use a complex combination of flat and structured features as well as the web, where we use relatively simple features with Wikipedia and prior sampling.

Furthermore, our approach manages to achieve the same level of end-to-end performance as the Within-Oracle run, while achieving a significantly better F-score. More pieces of evidence were actually classified correctly, though this was not reflected directly in the end-to-end score due to issues described above.

5.4 Error Analysis

We proceed to analyse parts of our end-to-end results to see what is causing errors in the temporal evidence classification task. We found several common problems.

Semantic inference Some problems had to do with the fact that several snippets require semantic inference. The fact that someone dies effectively ends any relationships that this person had. Another example is when someone marries someone (A marries C), and this beginning of relationships effectively

[1] As the Nil-Baseline is applied directly to the final tuples rather than the classification labels, there are is no F1 score for this run.

means the end of relationships for previous relations (A and B). A more complex method to deal with this type of semantic inference is needed, simple classification does not work so well. Here is an example:

> *Angela Merkel is married to Joachim Sauer, a professor of chemistry at Berlin's Humboldt University, since 1998. Divorced from Ulrich Merkel. No children.*

For this example the fact is that the time expression 1998 happens *after* with regard to the *spouse* relation between Angela Merkel and Ulrich Merkel.

Concise temporal representations Newspaper articles contain lots of temporal information in a concise way. For example in the form (X–Y). This implicit interval range is not expressed in a lexical context but rather with symbolic conventions. In several articles, the information encoded is almost tabular rather than expressed in explicitly. For example:

> *Elected as german chancellor Nov. 22, 2005. Chairwoman, christian democratic union, 2000-present. Chairwoman, christian democratic parliamentary group, 2002–2005.*

Complex time-inference BEFORE and AFTER are especially tricky to deal with because they require additional inference. Even if a passage contains the word *after*, the time expression linked to it would probably contain the *before* relation.

> *He was called up by the Army in the spring of 1944, after marrying bea silverman in 1943, and was sent to The Philippines.*

For the above example, 1943 happens *before* the "person joined the Army" event.

We observe quite a number of these cases on the evaluation data. Furthermore, the lack of context on some examples and evidence that is scattered around multiple sentences complicates the problem even more. Because of semantic and implicit evidence, temporal evidence classification remains a challenging task. In order to achieve a better absolute performance, collective classification/inference of evidence seems an interesting option.

6 Conclusion

We have presented a distant-supervision approach to temporal evidence classification. The main feature of our distant supervision approach is that we consider the prior distribution of classes in the target domain in order to better model the task. We show that our prior-informed distant supervision approach manages to outperform a purely supervised approach. Our method also achieves state-of-the-art performance on end-to-end temporal relation extraction with fewer and simpler features than previous work.

Our error analysis on the temporal evidence classification task revealed several issues that inform our future work aimed at further improving the performance on the subtask of temporal evidence classification, and the overall temporal relation extraction task. In particular, we intend to deal with the challenging aspect of semantic inference over relations found in the evidence passage. Another interesting direction that we aim to tackle is dealing with evidence that is scattered across multiple sentences.

Acknowledgements

This research was supported by the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreements nrs 288024 and 312827, the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, the Center for Creation, Content and Technology (CCCT), the QuaMerdes project funded by the CLARIN-nl program, the TROVe project funded by the CLARIAH program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project nr 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

References

Razvan C. Bunescu and Raymond J. Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, June.

Amev Burman, Arun Jayapal, Sathish Kannan, Ayman Kavilikatta, Madhu abd Alhelbawy, Leon Derczynski, and Robert Gauzauskas. 2011. USFD at KBP 2011: Entity linking, slot filling and temporal bounding. In *Proceedings of the TAC-KBP 2011 Workshop*. NIST.

Lorna Byrne and John Dunnion. 2011. UCD IIRG at tac 2011. In *Proceedings of the TAC-KBP 2011 Workshop*. NIST.

Guillermo Garrido, Anselmo Peñas, Bernardo Cabaleiro, and Álvaro Rodrigo. 2012. Temporally anchored relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 541–550, Stroudsburg, PA, USA. Association for Computational Linguistics.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC 2011 knowledge base population task. In *Proceedings of the TAC-KBP 2011 Workshop*. NIST.

Heng Ji, Taylor Cassidy, Qi Li, and Suzanne Tamang. 2013. Tackling representation, annotation and classification challenges for temporal knowledge base population. *Knowledge and Information Systems*, pages 1–36.

Qi Li, Javier Artiles, Taylor Cassidy, and Heng Ji. 2012. Combining flat and structured approaches for temporal slot filling or: how much to compress? In *Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'12, pages 194–205, Berlin, Heidelberg. Springer-Verlag.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL '09)*, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mattheiu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5*.

Ridho Reinanda, Daan Odijk, and Maarten de Rijke. 2013. Exploring entity associations over time. In *SIGIR 2013 Workshop on Time-aware Information Access*, August.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, ECML PKDD'10, pages 148–163, Berlin, Heidelberg. Springer-Verlag.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2003. Modeling missing data in distant supervision for information extraction. In *Transactions of the Association for Computational Linguistics*, TACL'13, pages 367–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Avirup Sil and Silviu Cucerzan. 2014. Temporal scoping of relational facts based on Wikipedia data. In *CoNLL: Conference on Natural Language Learning*.

Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, and Christopher D. Manning. 2011. Stanford's distantly-supervised slot-filling system. In *Proceedings of the TAC-KBP 2011 Workshop*. NIST.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Suzanne Tamang and Heng Ji. 2012. Relabeling distantly supervised training data for temporal knowledge base population. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pages 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. TempEval-3: Evaluating events, time expressions, and temporal relations. *CoRR*, abs/1206.5333.

Xu Wei, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2012. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 825–834, Stroudsburg, PA, USA. Association for Computational Linguistics.