Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5573–5579
Marseille, 20-25 June 2022

Enhanced Distant Supervision with State-Change Information for Relation Extraction

Jui Shah*,1, Dongxu Zhang*,1, Sam Brody[2], Andrew McCallum[1]
[1]University of Massachusetts Amherst, [2]Bloomberg
{jbshah, dongxuzhang, mccallum}@cs.umass.edu
sbrody18@bloomberg.net

Abstract

In this work, we introduce a method for enhancing distant supervision with state-change information for relation extraction. We provide a training dataset created via this process, along with manually-annotated development and test sets. We present an analysis of the curation process and data, and compare it to standard distant supervision. We demonstrate that the addition of state-change information reduces noise when used for static relation extraction, and can also be used to train a relation-extraction system that detects a change of state in relations.

Keywords: Corpus, Information Extraction, Linked Data, Weakly-supervised Learning, Relation Extraction

1. Introduction

Relation Extraction (RE) is an important task in natural language processing, both as an end goal (e.g., for knowledge-base population), and as a component in downstream applications such as search engines, recommendation systems, and question-answering systems. Most of the work in RE focuses on *static* relations, where the question of *when* the relation held is ignored. However, there has also been significant interest in temporal aspects of RE (see Section 2).

In this paper, we introduce a new task *State-change Relation Extraction*, which focuses specifically on the start and end of the relation, or *state-change* points. Detecting a change of state as it occurs is of value in and of itself, for example in an algorithmic trading system, that scans news feeds as they are published. And current knowledge bases (such as Wikidata) usually have time labels of relations to indicate the valid duration of them. In addition, change-of-state information can benefit systems focused on static relations, by reducing noise when using distant supervision (DS) (Mintz et al., 2009) since relationships mainly appear in news articles during the time they hold.

In this work, we create a dataset labeled with change-of-state information, present an analysis of the curation process and data, and compare it to classic distant supervision without temporal information. We use the temporally-linked dataset to train a relation-extraction system to detect change of state in four relations, and also experiment with using the change-of-state information to reduce noise in distant supervision for static relation extraction. We conclude with an analysis of our results and directions for future work.

Contributions of this paper mainly rely on three folds:

- We introduce a new task *State-change Relation Extraction* and provide a dataset with four relation types on general domain new corpus.
- We propose a simple yet effective enhancement on traditional distant supervision to capture the state-change relations via *time window alignment*.
- We also show that time window can de-noise the distant supervision signals for static relation extraction.

Our annotated data and the code for our experiments are available at `https://github.com/iesl/state-change-re`.

2. Related Work

Our work is based on “distant supervision”, a method proposed by Mintz et al. (2009) which uses relations from a knowledge-base (KB) to obtain unsupervised annotated data for relation extraction. In this method, every relation tuple (entity1, entity2, relation) in the KB is aligned with sentences in the corpus that mention both entities. Those sentences are considered as positive examples for the relation. A key weakness in this approach is the strong assumption that a given pair of entities are unlikely to co-occur in a sentence unless it expresses the relation of interest. This assumption may be violated in several ways, as seen in Section 3. Several methodologies have been proposed to overcome this weakness, including matrix factorization (Riedel et al., 2010), multi-instance multi-label learning (Surdeanu et al., 2012), neural-based selection (Zeng et al., 2015) and reinforcement learning (Feng et al., 2018), etc.

Yan et al. (2019) is one of the first works to investigate time information in distant supervision for relation extraction. Given a pair of related entities with a start and end time for the relation, they recognize time information from sentence context, and only align sentences when the time is within the specified time range. They then sort the list of aligned sentences and add time features based on the sorted rank into the text encoder. While their work focuses only on improving static RE,

* Equal contribution

ours also aims to predict changes in the relation state. Furthermore, we use the article publication timestamp, which allows our technique to be applied to a much wider range of sentences, including those that do not contain explicit time information in the text.

In Jiang et al. (2019) the authors use a template-based system to extract high confidence relation instances that serve in place of a knowledge-base. They use these relations to annotate data via distant supervision, but find that this approach leads to noisy annotation. To alleviate the issue, they propose a time aware strategy: by measuring the frequency of mentions of a given relation across time and multiple corpora, they find the popularity peak, and use only instances in temporal proximity to that point as training data. While the authors do not connect the inferred popularity peak with change-of-state events, we hypothesize that these often coincide. Their work is restricted to static relations, but complementary to ours in that they provide an unsupervised method for obtaining temporal information when a KB is not available but there is an abundance of relation-rich corpora.

While our work proposes the use of temporal information for RE, it is not directly connected to the area of research sometimes called "Temporal Relation Extraction". That body of work (e.g., The TimeBank corpus (Pustejovsky et al., 2003) and TempEval tasks (UzZaman et al., 2013)) is mainly focused on the understanding of temporal expressions and the order of events in a narrative, rather than structured relation extraction.

3. Distant Supervision Dataset with State-Change Information

As mentioned in Section 2, distant supervision is predicated on the assumption that if an entity pair are connected by a relation in the knowledge base, any sentence in a corpus in which the two entities appear conveys the existence of that relationship. In practice, many sentences violate this assumption, and this may be due to a variety of reasons: the entities may be incorrectly linked to the ones in the KB, the sentence may be indicating a different relation between the entities (which may or may not be present in the KB), or it may contain no relation at all. Even when the entities are correctly linked, and the sentence discusses the relation of interest, it may be referencing a past or future event, (e.g., "X and Y intend to wed next month"), making it unsuitable as a positive example for training.

To reduce the chance of error when associating a relation in the KB with sentences in the corpus, we introduce an enhanced version of distant supervision. This process makes use of temporal information in the form of the publication timestamp of news articles in our corpus, and the start and end dates associated with relations in the KB. By matching the article's publication date with the dates in the KB, we greatly increase the likelihood that the entity linking is correct, and that the sentence is discussing the relation of interest. In addition, this time-aligned data can be used to train a model to detect change-of-state in relations, which is not possible with standard distant-supervision.

3.1. Data Sources

We use WikiData (Vrandečić and Krötzsch, 2014) [1] as our knowledge base, and English Gigawords Fifth Edition (Parker et al., 2011) [2] as our text corpus.

WikiData is an open knowledge base that contains over 90 Million items, including over 9000 properties, and approximately 1.35 billion statements. Importantly, some statements include a time range, for example, ("Vladimir Putin", "head of state", "Russia", start time: 7 May 2000, end time: 7 May 2008). From this dataset, we select four properties (relation types): P26 (spouse), P35 (head of state), P463 (member of) and P54 (member of sports team), to demonstrate our enhanced DS method. These relations are widely attested in both the corpus and the WikiData knowledge base, indicating their importance, and making them good candidates for our experiments.

English Gigaword is a newswire text corpus which contains nearly ten million documents and covers seven news sources, including The New York Times, Agence France-Presse and Xinhua News Agency, collected by LDC between 1994-2010. Each document is labeled with a unique identifier that incorporates its date of publication.

To avoid overlap between training and test sets, we divided the dataset by years: 1994-2003 for training, 2004-2006 for development, and 2007-2010 for test.

3.2. Data Processing

We use Beautiful Soup[3] to tokenize each document from the newspaper archives. We then split it into sentences and identify the entities in each sentence using Spacy's `en_core_web_trf` NER model[4]. For each (ordered) pair of entity mentions in a sentence, we can construct a tuple (E_1, E_2, T), where E_1 and E_2 are the entity mentions, and T is the publication date of the document.

We use the WikiData API to extract statements of the form (e_1, e_2, r, t), where e_1, e_2 are subject and object entities, r is the relation type, and t is the start or end date.

Enhanced Distant Supervision Alignment: Given a time window specified by a pair of integers (w_1, w_2) (where the units are days), a sentence containing the tuple (E_1, E_2, T) will be linked with a WikiData statement (e_1, e_2, r, t) if $match(E_1, e_1) \wedge match(E_2, e_2) \wedge T \in [t+w_1, t+w_2)$. The entity linking function $match$ is a parameter. We employ a method proposed in previous work (Riedel et al., 2013): we first attempt to find

[1] We downloaded WikiData on 06-21-2021 from https://dumps.wikimedia.org/WikiDatawiki/latest/WikiDatawiki-latest-wb_items_per_site.sql.gz

[2] https://catalog.ldc.upenn.edu/LDC2011T07

[3] https://www.crummy.com/software/BeautifulSoup/

[4] https://spacy.io/models/en#en_core_web_trf

a full string match. If a string does not match any entity names, we apply TF-IDF similarity (Murty et al., 2018) with thresholding for better recall. Specifically, we collect all relevant WikiData entity names and create feature vectors with word level TF-IDF (vocabulary size = 100k, ngram $\in$ [1,3]) and character level TF-IDF (vocabulary size = 100k, ngram $\in$ [2,5]). We link entities if the cosine similarity of their feature vectors are greater than a threshold of 0.96.

Table 1 shows the number of aligned instances for each of the relations in the different datasets and window sizes.

3.3. Manual Annotation

To create our test and development sets, we first used the enhanced distant supervision method described above to generate annotation candidates. We then randomly selected 10 candidates from each time window interval as listed in Table 1[5]. The first three authors of this paper manually annotated these instances, for static and change-of-state, following the guidelines in the Appendix. All annotators reviewed all sentences. The datasets include the annotation in the form of a confidence score: the fraction of annotators who labeled the instance as positive.

This process provided a challenging test set, since it did not include any "easy" negatives - all sentences contained entity pairs which participated in the target relation at some point in time (according to the DS signal).

3.4. Discussion

Table 2 shows the proportion of positive instances in the development and test sets. We calculated Kappa (Fleiss, 1971) to show the agreement among annotators. For static relations, around a third of the instances were labeled as negative, indicating the noise inherent in using distantly-supervised data, and provide counterexamples to the underlying assumption of DS, that two entities participating in a relation are unlikely to appear in a sentence whose semantics do not express that relation. Some common examples we saw in our dataset include mentions of a head-of-state and their country (e.g., Putin and Russia) in sentences that does not express that relation, or sentences listing the cast of a movie, in which two of the actors were married (possibly after the time the sentence was written).

For temporal relations, a much smaller portion of the instances were labeled positive, demonstrating that sentences indicating a change of state are much rarer than those indicating the static relation. For this reason, relation P26 (spouse) is the only one for which we consider the "end" temporal relation: all others had fewer than 10 positive instances in the development or test set, and were therefore excluded. The rarity of such sentences suggests that obtaining sufficient positive training examples for detecting change-of-state would be difficult using manual approaches or simple heuristics.

Change-of-state relations add an additional challenge, since they require a clear definition of when a relation starts or ends, and this may be a function of the intended use case. For example, the spousal relation could be considered ended when the couple publicly separate, when the divorce proceedings are initiated, or when they have legally concluded. An athlete may be considered a member of the team from the time they signed a contract for the upcoming season, or only when the season has started. This added complexity was evident in the decreased inter-annotator agreement for the temporal relations compared to the static ones (as seen in Table 2), despite the establishment of specific guidelines which attempted to consider such edge cases.

4. Experimental Setup

Model: We use the OpenNRE framework[6] (Han et al., 2019) to conduct our experiments. Specifically, we use the supervised sentence-level relation extraction method with bert-base-uncased as the encoder, and the default relation representation (entity representation concatenation). The representation is fed to a sigmoid classification head consisting of one dropout and one fully connected layer. We use binary cross entropy with logit loss as the loss function.

Training: We train the model for 10 epochs using the following hyperparameters: batch size of 64, learning rate 0.0001, max sequence length of 128. We did not employ early stopping, and used the development set only for calculating the optimal threshold for classification.

Data: We regard a test or development instance as positive if at least one annotator labeled it as such (confidence $>$ 0.33). All our experiments used a single set of 10k negative training examples randomly sampled from sentences containing two entities which did not match any instances of our target relation types in WikiData. The selection of positive training examples was the driving parameter of the experiments.

Experiments: We conducted two sets of experiments corresponding to the two use cases we proposed in Section 3: static RE, and state-change RE. In the static RE use case, the enhanced DS method is used to improve the quality of the linking, and reduce noise in our training dataset. We compare using our enhanced DS with various window sizes to a baseline where different amounts of training data obtained via standard DS are used. In the state-change RE use case, we are interested in detecting sentences that indicate a change of state in the relation. For this use case, standard DS is not directly applicable.

[5] For the spouse end relation, we did not sample instances from the 'no window' interval, to avoid duplication with spouse start.

[6] https://github.com/thunlp/OpenNRE

Relation	P26 start	P26 end	P35 start	P54 start	P463 start
no window	37k / 14k / 22k		98k / 43k / 61k	66k / 34k / 47k	298k / 103k / 140k
[-300, -100)	272 / 127 / 87	389 / 290 / 213	881 / 287 / 678	195 / 245 / 242	4,264 / 265 / 430
[-100, -30)	127 / 53 / 102	142 / 132 / 131	1,289 / 1,065 / 1,715	138 / 152 / 106	1,861 / 625 / 141
[-30, -10)	42 / 68 / 151	34 / 33 / 28	436 / 400 / 466	29 / 12 / 21	488 / 264 / 45
[-10, 10)	403 / 154 / 359	181 / 118 / 301	1,047 / 824 / 1,361	42 / 34 / 90	1,663 / 1,237 / 438
[10, 30)	102 / 33 / 32	56 / 28 / 41	462 / 333 / 612	74 / 48 / 131	538 / 144 / 39
[30, 100)	420 / 44 / 124	201 / 54 / 213	1,175 / 593 / 1,194	418 / 408 / 419	1,985 / 374 / 203
[100, 300)	819 / 246 / 426	456 / 204 / 366	2,685 / 950 / 3,312	1,865 / 1,558 / 2,030	2,673 / 818 / 218

Table 1: Number of distantly aligned sentences in the training/development/test portions of the Gigaword corpus, for each relation and time window (in days) . (P26: spouse; P35: head of state; P54: member of sports team; P463: member of.). For the classic DS setting (no window), start and end information is irrelevant.

Relation	Dev.	Test
Static Relation. $\kappa = 0.755$		
P26	86 / 140 (61.4%)	98 / 150 (65.3%)
P35	14 / 80 (17.5%)	51 / 80 (63.8%)
P54	62 / 80 (77.5%)	57 / 80 (71.3%)
P463	53 / 80 (66.3%)	58 / 80 (72.5%)
Change of State. $\kappa = 0.676$		
P26 start	21 / 80 (26.3%)	14 / 80 (17.5%)
P26 end	19 / 60 (31.7%)	34 / 70 (48.6%)
P35 start	14 / 80 (17.5%)	13 / 80 (16.3%)
P54 start	25 / 80 (31.3%)	25 / 80 (31.3%)
P463 start	13 / 80 (16.3%)	11 / 80 (13.8%)

Table 2: Proportion of positive instances in the dev and test sets. Kappa κ among three annotators are shown in the table. (P26: spouse; P35: head of state; P54: member of sports team; P463: member of.)

Time window	Static Rel.	State-Change
no window	72.2%	2.8%
[-300, -100)	33.0%	10.0%
[-100, -30)	48.0%	16.7%
[-30, -10)	55.4%	22.8%
[-10, 10)	**89.1%**	**72.3%**
[10, 30)	87.8%	35.7%
[30, 100)	77.0%	10.0%
[100, 300)	83.0%	19.8%

Table 3: Accuracy of the distant supervision heuristic using different time windows, with human annotation as ground truth. The left column (**Static Rel.**) indicates the quality of signals for static relations, and the right column (**State-Change**) indicates the quality of signals for change-of-state relations. Results are calculated across all relation types in the dev and test set.

5. Results

5.1. Quality of Time Windows

In this section, we directly calculate the accuracy of distant supervision signals on our curated test set as a sanity check of whether a time window can provide cleaner signals for both static and temporal relation extraction.

Table 3 shows that for both static and change-of-state relation extraction, the time window can indeed provide cleaner signals. When targeting change-of-state, all the time windows we applied show better quality than standard distant supervision (without a time window). For static relation extraction, sentences posted after the relation start date provide a much better signal than ones posted before, validating our hypothesis that temporal information can be used to reduce noise in distant supervision.

The numbers in the table also show that the quality of the signal for the change-of-state RE is much lower than for static RE, even when restricting to a time window. This is a further indication of the difficulty of the change-of-state RE task, as mentioned in Section 3.4.

5.2. Static Relation Extraction

Table 4 shows the performance of the models trained on our enhanced DS data (bottom) as compared to baseline models trained with standard DS data (top). Note that for the spouse relation (P26), we trained separate models using the start or end signal, but evaluate on a combined test set containing 150 static annotations for the spouse relation (see Table 2).

Our results show several trends. On the baseline side, the best results are usually obtained on the middle setting of 5,000 positive training instances. Adding more positive instances increases noise and reduces performance. On the experiment side, the best window choice varies across relations. For spouse relations, using enhanced DS with window $[-30, 30)$ strongly outperforms standard DS. For the other relations, the best setting for classic DS performs better. One possible explanation is that the assumption underlying distant supervision is more likely to hold true for these relations than for others, so the benefit of more training data outweighs the noise reduction provided by the change-of-state signal.

# +ve	P26s	P26e	P35s	P463s	P54s	Macro
1,000	66.7	62.6	63.2	**67.8**	77.5	67.6
5,000	55.3	62.6	**79.3**	66.7	**84.1**	**69.6**
10,000	**77.4**	62.6	69.1	64.0	61.5	66.9

Window	P26s	P26e	P35s	P463s	P54s	Macro
[-10,10)	65.5	**83.0**	**70.9**	**63.4**	63.2	**69.2**
[-30,30)	**79.4**	81.2	66.0	51.7	59.1	67.5
[-100,100)	74.6	79.6	53.8	54.4	**74.5**	67.4
[-300,300)	76.9	80.3	18.5	42.2	59.6	55.5

Table 4: % F-1 scores for static relations using standard (top) and enhanced (bottom) distant supervision. For standard DS, results for different numbers of positive instances are shown. For enhanced DS, the window used for positives and hard negatives is shown. In both cases, 10k sampled negatives were used in training. Here, P26: Spouse, P35: Head of state, P463: Member of, P54: Member of sports team, s/e: start/end.

Window	P26s	P26e	P35s	P463s	P54s	Macro
[-10,10)	38.6	**61.2**	**52.9**	**24.0**	**66.7**	**48.7**
[-30,30)	**61.1**	55.9	41.4	18.2	65.4	48.4
[-100,100)	41.9	43.9	35.7	19.1	58.8	39.9
[-300,300)	48.3	46.8	37.0	21.1	47.1	40.1

Table 5: % F-1 scores for change-of-state in relations using enhanced distant supervision with various window sizes. Here, P26: Spouse, P35: Head of state, P463: Member of, P54: Member of sports team, s/e: start/end.

5.3. State-Change Relation Extraction

Table 5 presents the results of our experiments using enhanced DS for detecting change-of-state in relations. As discussed previously (see Section 3.3) this is a difficult task, where human annotators also struggle in some cases, and this is reflected in the significantly lower scores compared to the corresponding static relations (Table 4). With the exception of spouse-start, the narrowest window around the time of state change yields the best results. In all cases, expanding the window beyond $[-30, 30)$ reduces performance. These results confirm that the enhanced DS method we proposed can provide suitable data for training a model that can detect changes in relation state.

6. Summary

In this work, we introduce an enhanced version of distant supervision for relation extraction, which uses change-of-state information to provide tighter linking between the knowledge-base relations and sentences in the corpus. The data obtained via this method can be used to reduce noise when training a standard static relation extraction model, or to train models that specifically detect changes in relationship state. We construct a training dataset for four relations using our enhanced distant-supervision technique, and manually annotate corresponding development and test sets. We experiment with using this dataset on both the static and change-detection scenarios and present our results.

6.1. Directions for Future Work

In this work, we made use of a corpus and knowledge base that are easily accessible and widely used in the academic literature, for easy replication and comparison to previous work. The general news domain of The New York Times corpus made it likely to contain a wide range of relations, but also skewed those mentions towards popular “head” entities. Similar biases are likely present in the WikiData knowledge-base. In future work, we would like to focus on specific relations of interest, and choose the corpus and knowledge-base accordingly. For example, for corporate action relations (e.g. mergers and acquisitions) a financial corpus such as The Wall Street Journal may give wider coverage.

We also made several decisions regarding relative time expressions. We disregarded statements about future events, even when they specified concrete dates and had high probability of occurring, and we did not attempt to resolve relative time expressions to exact dates, preferring to use time windows instead. These decisions were intended to simplify the experimental setup and learning process, but a more thorough treatment of relative time expressions is likely to yield performance improvements.

Acknowledgments

This work is based upon work supported in part by the Center for Data Science and the Center for Intelligent Information Retrieval, and in part by the National Science Foundation under Grant Nos. 1763618 and 1514053, and in part by the Chan Zuckerberg Initiative under the project “Scientific Knowledge Base Construction”. Some of the work reported here was performed using high performance computing equipment obtained under a grant from the Collaborative R&D Fund managed by the Massachusetts Technology Collaborative. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

7. Bibliographical References

Feng, J., Huang, M., Zhao, L., Yang, Y., and Zhu, X. (2018). Reinforcement learning for relation classification from noisy data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., and Sun, M. (2019). OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*, pages 169–174.

Jiang, T., Zhao, S., Liu, J., ge Yao, J., Liu, M., Qin, B., Liu, T., and Lin, C.-Y. (2019). Towards time-aware distant supervision for relation extraction. *ArXiv*, abs/1903.03289.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (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*, pages 1003–1011, Suntec, Singapore, August. Association for Computational Linguistics.

Murty, S., Verga, P., Vilnis, L., Radovanovic, I., and McCallum, A. (2018). Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 97–109.

Riedel, S., Yao, L., and McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, September.

Riedel, S., Yao, L., McCallum, A., and Marlin, B. M. (2013). Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.

Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (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*, pages 455–465, Jeju Island, Korea, July. Association for Computational Linguistics.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Yan, J., He, L., Huang, R., Li, J., and Liu, Y. (2019). Relation extraction with temporal reasoning based on memory augmented distant supervision. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1019–1030, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zeng, D., Liu, K., Chen, Y., and Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal, September. Association for Computational Linguistics.

8. Language Resource References

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword fifth edition, 2011. *Linguistic Data Consortium, Philadelphia, PA, USA*.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003). The timebank corpus. *Proceedings of Corpus Linguistics*, 01.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Appendix: Annotation Guidelines

8.1. General Guidelines for State-Change

If a sentence does not contain a time expression, only label it as true if it contains an expression indicating a change-of-state of the relation (For example, "(got) divorced" indicates a change of state, "ex-wife" does not). If a sentence contains relative time information (for example, "just now", "yesterday", "last week", etc.), label it as true if the point in time happened within the last 30 days. A prediction of future relationship should always be regarded as false. If a sentence contains absolute time information (for example, "on September 5th"), then label it as true if the timestamp of the sentence is within the last 30 days (e.g., the time information "on September 5th" appeared in a document with timestamp "2007-9-25").

8.2. Individual Relations

8.2.1. "spouse"

Only consider as positive if marriage is implied (wife, husband, spouse, etc.). "Couple", "fiance", or shared children do not imply marriage. Start is the date of the wedding. End is the date of the announcement of divorce or split-up, or beginning of divorce proceedings.

8.2.2. "head of state"

Only includes "president", excludes "prime minister", etc. and should only include officially recognized roles. A person is not considered the head of state until they have been sworn in or otherwise started to function as head of state. Winning the election, being "president-elect", or "incoming president" should be marked negative. Acting head of state should be considered positive. "Outgoing president" is still president until they no longer hold the position (resign or are replaced).

Future intents should be labeled negative. “New president” does not indicate time, and should only be labeled positive if a separate time expression is present (absolute or relative).

8.2.3. “member of sports team”

Only true when the subject entity officially becomes the member of the object, for example, when a contract is signed, or when someone has announced the member “(has) joined the team”. Future commitments to join should be annotated as negative, but signing a contract to play next season should be annotated positive since the person has officially joined the team. For end dates, only an official announcement (from player or team) should be labeled positive, and only if it does not refer to the future.

8.2.4. “member of”

Only true when the subject entity officially becomes the member of the object, for example, when a contract is signed, or when someone has announced the member is hired. Future commitments to join should be annotated as negative (e.g., “intends to join Y next year”)