Temporal information extraction from clinical text

Julien Tourille, Olivier Ferret, Xavier Tannier, Aurélie Névéol

Temporal information extraction from clinical text

Julien Tourille
LIMSI, CNRS
Univ. Paris-Sud
Université Paris-Saclay
julien.tourille@limsi.fr

Olivier Ferret
CEA, LIST,
Gif-sur-Yvette,
F-91191 France.
olivier.ferret@cea.fr

Xavier Tannier
LIMSI, CNRS
Univ. Paris-Sud
Université Paris-Saclay
xavier.tannier@limsi.fr

Aurélie Névéol
LIMSI, CNRS
Université Paris-Saclay
aurelie.neveol@limsi.fr

Abstract

In this paper, we present a method for temporal relation extraction from clinical narratives in French and in English. We experiment on two comparable corpora, the MERLOT corpus for French and the THYME corpus for English, and show that a common approach can be used for both languages.

1 Introduction

Temporal information extraction from electronic health records has become a subject of interest, driven by the need for medical staff to access medical information from a temporal perspective (Hirsch et al., 2015). Diagnostic and treatment could be indeed enhanced by reviewing patient history synthetically in the order in which medical events occurred. However, most of this temporal information remains locked within unstructured texts and requires the development of NLP methods in order to be accessed.

In this paper, we focus on the extraction of temporal relations between medical events (`EVENT`), temporal expressions (`TIMEX3`) and document creation time (DCT). More specifically, we address intra-sentence narrative container relation identification between medical events and/or temporal expressions (CR task, for **C**ontainer **R**elation) and DCT relation identification between medical events and documents (DR task, for **D**ocument creation time **R**elation).

In the DR task, the objective is to temporally locate `EVENT` entities according to the Document Creation Time of the document in which they occur. Possible tags are *Before*, *Before-Overlap*, *Overlap* and *After*.

In the CR task, the objective is to identify temporal inclusion relations between pairs of entities (`EVENT` and/or `TIMEX3`) formalized as narrative container relations following Pustejovsky and Stubbs (2011).

In this context, we build on Tourille et al. (2016) and show how this type of model can be applied for extracting temporal relations from clinical texts similarly in two languages. We experimented more specifically on two corpora: the THYME corpus (Styler IV et al., 2014), a corpus of de-identified clinical notes in English from the Mayo Clinic and the MERLOT corpus (Campillos et al., to appear), a comparable corpus in French from a group of French hospitals.

2 Related Work

Temporal information extraction from clinical texts has been the topic of several shared tasks over the past few years.

The i2b2 Challenge for Clinical Records (Sun et al., 2013) offered to work on events, temporal expressions and temporal relation extraction. Participants were challenged to detect clinically relevant events and time expressions and link them with a temporal relation.

SemEval has been offering the Clinical TempE-

val task related to the topic for the past two years (Bethard et al., 2015; Bethard et al., 2016). Its first track focused on extracting clinical events and temporal expressions, while its second track included DR and CR tasks. Different approaches were implemented by the teams, among which SVM classifiers (Lee et al., 2016; Tourille et al., 2016; Cohan et al., 2016; AAl Abdulsalam et al., 2016) and CRF approaches (Caselli and Morante, 2016; AAl Abdulsalam et al., 2016) for the DR task, and CRF, Convolutional neural networks (Chikka, 2016) and SVM classifiers (Tourille et al., 2016; Lee et al., 2016; AAl Abdulsalam et al., 2016) for the CR task.

3 Corpus Presentation

The MERLOT corpus is composed of clinical documents written in French from a Gastroenterology, Hepatology and Nutrition department. These documents have been de-identified (Grouin and Névéol, 2014) and annotated with entities, temporal expressions and relations (Deléger et al., 2014). The THYME corpus is a collection of clinical texts written in English from a cancer department that have been released during the Clinical TempEval campaigns. This corpus contains documents annotated with medical events and temporal expressions as well as container relations.

The definition of a medical event is slightly different in each corpus. According to the annotation guidelines of the THYME corpus, a medical event is anything that could be of interest on the patient's clinical timeline. It could be for instance a *medical procedure*, a *disease* or a *diagnosis*. There are five attributes given to each event: *Contextual Modality* (Actual, Hypothetical, Hedged or Generic), *Degree* (Most, Little or N/A), *Polarity* (Pos or Neg), *Type* (Aspectual, Evidential or N/A) and *DocTimeRel* (Before, Before-Overlap, Overlap and After). Concerning the temporal expressions, a *Class* attribute is given to each of them: Date, Time, Duration, Quantifier, Pre-PostExp or Set.

For the French corpus, medical events are described according to UMLS® (Unified Medical Language System) Semantic Groups and Semantic Types. Several categories are considered as events: *disorder*, *sign or symptom*, *medical procedure*, *chemical and drugs*, *concept or idea* and *biological process or function*. Events carry only one *DocTime* attribute (Before, Before-Overlap, Overlap or After). Similarly to the THYME corpus, temporal expressions within the French corpus are given a class among: Date, Time, Duration or Frequency.

Narrative containers (Pustejovsky and Stubbs, 2011) can be apprehended as temporal buckets in which several events may be included. These containers are anchored by temporal expressions, medical events or other concepts. Styler IV et al. (2014) argue that the use of narrative containers instead of classical temporal relations (Allen, 1983) yields better annotation while keeping most of the useful temporal information intact. The concept of narrative container is illustrated in Figure 1 and described further in Pustejovsky and Stubbs (2011).

Figure 1: Examples of intra-sentence narrative container relations.

The French corpus does not explicitly cover container relations. However, we consider that *During* relations are equivalent to *Contains* relations. In addition, we also considered that *Reveals* and *Conducted* relations imply *Contains* relations. Furthermore, the corpus does not cover inter-sentence relations (relations that can spread over multiple sentences). We focus in this paper on intra-sentence container relations (relations that are embedded within the same sentence) and we will refer to them as `CONTAINS` relations in the rest of this paper.

Descriptive statistics of the two corpora are provided in Table 1.

4 Model Description

In our model, we consider both DR and CR tasks as supervised classification problems. Concerning the DR task, each medical event is classified into one category among *Before*, *Before-Overlap*, *Overlap* and *After*. The number of document creation time relations per class for both corpora is presented at table 3. For the CR task, we are dealing with a binary classification problem for each pair of `EVENT` and/or `TIMEX3`. However, considering all pairs of entities within a sentence would give us an unbalanced data set with a very large amount of negative examples. Thus, to reduce the number of candidate pairs, we transformed the 2-

	THYME	MERLOT
Tokens	501,156	179,200
EVENT[a]	DR 78,901 CR 64,650	18,127
TIMEX3[a]	DR 7,863 CR 7,708	3,940
CONTAINS	17,444	4,295

[a] Not all documents are annotated with container relations. We present separate count of EVENT and TIMEX3 for each task CR and DR.

Table 1: MERLOT (fr) and THYME (en) corpora – Descriptive Statistics.

Feature	DR	Container	CR
Entity type	✓	✓	✓
Entity form	✓	✓	✓
Entity attributes	✓	✓	✓
Entity position (within the document)	✓	✓	✓
Container model output			✓
Document Type[a]	✓	✓	✓
Contextual entity forms	✓	✓	✓
Contextual entity types	✓	✓	✓
Contextual entity attributes	✓	✓	✓
Container model output for contextual entities			✓
PoS tag of the sentence verbs	✓	✓	
Contextual token forms (unigrams)	✓	✓	
Contextual token PoS tags (unigrams)	✓	✓	
Contextual token forms (bigrams)[b]	✓	✓	
Contextual token PoS tags (bigrams)[b]	✓	✓	

[a] Information available only for the MERLOT corpus.
[b] Only when using plain lexical forms.

Table 2: Features used by our classifiers.

category problem (*contains* or *no-relation*) into a 3-category problem (*contains*, *is-contained*, or *no-relation*). In other words, instead of considering all permutations of entities within a sentence, we consider all combinations of entities from left to right, changing when necessary the *contains* relations into *is-contained* relations. Moreover, this transformation solves the problem of possible contradictory predictions. If we were to consider all pairs of entities within a sentence, we could have the situation where the prediction of our classifier implies that two entities contain each other (A contains B and B contains A). By considering all combinations instead of all permutations, the problem will never occur during the prediction phase. However, our system does not handle temporal closure, and conflicts could still appear at sentence level (X contains Y, X is contained by Z, Y contains Z).

	THYME (en)	MERLOT (fr)
Before	29,170	1,936
Bef./Over.	4,240	2,643
Overlap	37,091	12,211
After	8,400	1,337

Table 3: MERLOT (fr) and THYME (en) corpora - Document Creation Time relation repartition.

Furthermore, some entities are more likely to be the anchor of narrative containers. For instance, temporal expressions are, by nature, potential anchors and may contain other temporal expressions and/or medical events. This is also the case for some medical events. For instance, a *surgical operation* may contain other events such as *bleeding* or *suturing* whereas it will not be the same with the two latter in most cases. Following this observation, we have built a model to classify entities as being potential container anchors or not (CONTAINER classifier). This classifier obtains a high performance. We use its output as feature for our CONTAINS relation classifier.

4.1 Preprocessing and Feature Extraction

The THYME corpus has been preprocessed using cTAKES (Savova et al., 2010), an open-source natural language processing system for extraction of information from electronic health records. We extracted several features from the output of cTAKES: sentences boundaries, tokens, part-of-speech (PoS) tags, token types and semantic types of the entities that have been recognized by cTAKES and that have a span overlap with at least one EVENT entity of the THYME corpus.

Concerning the MERLOT corpus, no specific pipeline exists for French medical texts; we thus used Stanford CoreNLP system (Manning et al., 2014) to segment and tokenize the text. We also extracted PoS tags. As the corpus already provides a type for each EVENT, there is no need for detecting other medical information.

For both DR and CR tasks, we used a combination of structural, lexical and contextual features yielded from the corpora and the preprocessing steps. These features are presented in Table 2.

4.2 Lexical Feature Representation

We implemented two strategies to represent the lexical features in both DR and CR tasks. In the

Corpus	DCT		CONTAINER		CONTAINS		CONTAINS without CONTAINER	
	Plain	W2V	Plain	W2V	Plain	W2V	Plain	W2V
MERLOT (fr)	0.830 (0.008)	0.785 (0.006)	0.837 (0.004)	0.776 (0.014)	0.827 (0.007)	0.799 (0.012)	0.724 (0.011)	0.670 (0.016)
THYME (en)	0.868 (0.002)	0.797 (0.006)	0.760 (0.007)	0.678 (0.031)	0.751 (0.003)	0.702 (0.013)	0.589 (0.006)	0.468 (0.018)

(a) Cross-validation results over the training corpus for all tasks. We report F1-measure for CONTAINER and CONTAINS tasks and accuracy for DCT task. We also report standard deviation for all models.

	MERLOT (fr)			THYME (en)		
	P	R	F1	P	R	F1
baseline	0.67	0.67	0.67	0.47	0.47	0.47
bef./over.	0.68	0.69	0.69	0.73	0.60	0.66
before	0.81	0.60	0.69	0.88	0.88	0.88
after	0.79	0.69	0.73	0.84	0.84	0.84
overlap	0.88	0.92	0.90	0.88	0.90	0.89
micro-average	0.83	0.84	0.83	0.87	0.87	0.87

(b) DR task results over the test corpus. We report precision (P), recall (R) and F1-Measure (F1) for all relation types.

	MERLOT (fr)			THYME (en)		
	P	R	F1	P	R	F1
baseline	0.43	0.15	0.22	0.55	0.06	0.11
no-relation	0.99	1.00	0.99	0.96	0.98	0.97
contains	0.75	0.57	0.65	0.61	0.47	0.53
micro-average	0.98	0.98	0.98	0.93	0.94	0.93

(c) CR task results over the test corpus. We report precision (P), recall (R) and F1-Measure (F1) for all relation types.

Table 4: Experimentation results.

first one, we used the plain forms of the different lexical attributes we mentioned in the previous section. In the second strategy, we substituted the lexical forms with word embeddings. For English, these embeddings have been computed on the Mimic 3 corpus (Saeed et al., 2011). Concerning the French language, we used the whole collection of raw clinical documents from which the MERLOT corpus has been built. In both cases, we computed[1] the word embeddings using the word2vec (Mikolov et al., 2013) implementation of gensim (Řehůřek and Sojka, 2010). We used the max of the vectors for multi-word units. Lexical contexts are thus represented by 200-dimensional vectors. When several contexts are considered, e.g. right and left, several vectors are used.

5 Experimentation

We divided randomly the two corpora into train and test set following the ratio 80/20. We performed hyper-parameter optimization using a Tree-structured Parzen Estimator approach (Bergstra et al., 2011), as implemented in the library *hyperopt* (Bergstra et al., 2013), to select the hyper-parameter C of a Linear Support Vector Machine, the lookup window around entities and the percentile of features to keep. For the latter we used the ANOVA F-value as selection criterion. We used the SVM implementation provided within *Scikit-learn* (Pedregosa et al., 2011). In each case, we performed a 5-fold cross-validation. For the container classifier and contains relation classifier, we used the F1-Measure as performance evaluation measure. Concerning the DCT classifier, we used the accuracy.

[1]Parameters used during computation: algorithm = CBOW; min-count = 5; vector size = 200; window = 10.

6 Results and Discussion

Cross-validation results are presented in Table 4a. DR and CR tasks results are presented respectively in Table 4b and Table 4c. For both tasks, we present a baseline performance. For the DR task, the baseline predicts the majority class (*overlap*) for all `EVENT` entities. For the CR task, the baseline predicts that all `EVENT` entities are contained by the closest `TIMEX3` entity within the sentence in which they occur.

Concerning the DR task, there is a gap of 0.04 in performance between the French (0.83) and English (0.87) corpora. We notice that results per category are not homogeneous in both cases. Concerning the MERLOT corpus, the score obtained for the category *Overlap* is better (0.90) than the score obtained for *Before-Overlap* (0.69), *Before* (0.69) and *After* (0.73). Concerning the THYME corpus, the performance for the category *Before-Overlap* (0.66) is clearly detached from the

others which are grouped around 0.85 (0.88 for *Before*, 0.84 for *After* and 0.89 for *Overlap*). This may be due to the distribution of categories among the corpora. Typically, the performance is lower for the categories where we have a lower number of training examples (*Before-Overlap* for the THYME corpus and categories other than *Overlap* for the MERLOT corpus).

Concerning the CR task, results are separated by a 10 percent gap (0.65 for the MERLOT corpus and 0.53 for the THYME corpus). Results obtained for the THYME corpus are coherent with those obtained by Tourille et al. (2016) on the Clinical TempEval 2016 evaluation corpus[2]. We increased the recall value in comparison to their results (from 0.436 to 0.47) but this measure is still the main point to improve.

More globally, the best results of the Clinical TempEval shared task were 0.843 (accuracy) for the DR task and 0.573 (F1-Measure) for the CR task, which are comparable to our results (0.87 for the DR task and 0.53 for the CR task).

Table 4a also indicates that replacing lexical forms by word embeddings seems to have a negative impact on performance in every case.

As for the difference of performance according to the language, several parameters can affect the results. First, the sizes of the corpora are not comparable. The THYME corpus is bigger and has more annotations than the MERLOT corpus. Second, the quality of annotations is more formalized and refined for the MERLOT corpus. This difference can influence the performance, especially for the CR task. Third, the lack of specialized clinical resources for French can negatively influence the performance of all classifiers.

Concerning the quality of annotations, it has to be pointed out that inter-annotator agreement (IAA) for temporal relation is low to moderate: in MERLOT, IAA measured on a subset of the corpu s is 0.55 for *During* relations, 0.32 for *Conducted* relations and 0.64 for *Reveals* relations. In Thyme, IAA for *Contains* relation is 0.56. The inter-annotator agreement is comparable in both languages, and suggests that temporal relation extraction is a difficult task even for humans to perform.

[2]Similarly to our evaluation corpus for English, the Clinical TempEval 2016 evaluation corpus was extracted from the THYME corpus but the two corpora are different.

7 Conclusion and Perspectives

In this article, we have presented a work focusing on the extraction of temporal relations between medical events, temporal expressions and document creation time from clinical notes. This work, based on a feature engineering approach, obtained competitive results with the current state-of-the-art and led to two main conclusions. First, the use of word embeddings in place of lexical features tends to degrade performance. Second, our feature engineering approach can be applied with comparable results to two different languages, English and French in our case.

To follow-up with the first conclusion, we would like to test a more integrated approach for using embeddings, either by turning all features into embeddings as in Yang and Eisenstein (2015) or by adopting a neural network architecture as in Chikka (2016).

Acknowledgements

The authors thank the Biomedical Informatics Department at the Rouen University Hospital for providing access to the LERUDI corpus for this work. This work was supported in part by the French National Agency for Research under grant CABeRneT ANR-13-JS02-0009-01 and by Labex Digicosme, operated by the Foundation for Scientific Cooperation (FSC) Paris-Saclay, under grant CÔT.

References

Abdulrahman AAl Abdulsalam, Sumithra Velupillai, and Stephane Meystre. 2016. UtahBMI at SemEval-2016 Task 12: Extracting Temporal Information from Clinical Text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1256–1262, San Diego, California, June. Association for Computational Linguistics.

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc.

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vi-

sion Architectures. In *Proceedings of The 30th International Conference on Machine Learning*, pages 115–123.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado, June. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California, June. Association for Computational Linguistics.

Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. to appear. A French clinical corpus with comprehensive semantic annotations: Development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation*.

Tommaso Caselli and Roser Morante. 2016. VUA-CLTL at SemEval 2016 Task 12: A CRF Pipeline to Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1241–1247, San Diego, California, June. Association for Computational Linguistics.

Veera Raghavendra Chikka. 2016. CDE-IIITH at SemEval-2016 Task 12: Extraction of Temporal Information from Clinical documents using Machine Learning techniques. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1237–1240, San Diego, California, June. Association for Computational Linguistics.

Arman Cohan, Kevin Meurer, and Nazli Goharian. 2016. GUIR at SemEval-2016 task 12: Temporal Information Processing for Clinical Narratives. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1248–1255, San Diego, California, June. Association for Computational Linguistics.

Louise Deléger, Cyril Grouin, Anne-Laure Ligozat, Pierre Zweigenbaum, and Aurélie Névéol. 2014. Annotation of specialized corpora using a comprehensive entity and relation scheme. In *Proceedings of Language and Resource Evaluation Conference, LREC 2014*, pages 1267–1274.

Cyril Grouin and Aurélie Névéol. 2014. De-Identification of Clinical Notes in French: towards a Protocol for Reference Corpus Developpement. *Journal of Biomedical Informatics*, 50:151–61, Aug.

Jamie S. Hirsch, Jessica S. Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. 2015. HARVEST, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274.

Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1292–1297, San Diego, California, June. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

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, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James Pustejovsky and Amber Stubbs. 2011. Increasing Informativeness in Temporal Annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39:952–960, May.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010.

Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

William Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016. LIMSI-COT at SemEval-2016 Task 12: Temporal relation identification using a pipeline of classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1136–1142, San Diego, California, June. Association for Computational Linguistics.

Yi Yang and Jacob Eisenstein. 2015. Unsupervised Multi-Domain Adaptation with Feature Embeddings. In *2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2015)*, pages 672–682, Denver, Colorado, May–June.