Automatic Extraction of Time Expressions Across Domains in French Narratives

Mike Donald Tapi Nzali
Université de Montpellier
LIRMM, I3M
Montpellier, France

Xavier Tannier
LIMSI-CNRS UPR 3251
Univ. Paris-Sud
91403 Orsay, France
xtannier@limsi.fr

Aurélie Névéol
LIMSI-CNRS UPR 3251
Rue John von Neuman
91403 Orsay, France
neveol@limsi.fr

Abstract

The prevalence of temporal references across all types of natural language utterances makes temporal analysis a key issue in Natural Language Processing. This work adresses three research questions: 1/is temporal expression recognition specific to a particular domain? 2/if so, can we characterize domain specificity? and 3/how can subdomain specificity be integrated in a single tool for unified temporal expression extraction? Herein, we assess temporal expression recognition from documents written in French covering three domains. We present a new corpus of clinical narratives annotated for temporal expressions, and also use existing corpora in the newswire and historical domains. We show that temporal expressions can be extracted with high performance across domains (best F-measure 0.96 obtained with a CRF model on clinical narratives). We argue that domain adaptation for the extraction of temporal expressions can be done with limited efforts and should cover pre-processing as well as temporal specific tasks.

1 Introduction

References to phenomena occurring in the world and their temporal characterization can be found in natural language utterances across domains, genres and languages. Temporal analysis is a key issue in natural language processing that has been receiving increasing attention in recent years. Many efforts in this direction focused on newswire text in English. The focus on this language and domain was in part guided by the availability of the TimeBank corpus (Pustejovsky et al., 2003) used in evaluation campaigns such as TempEval (Verhagen et al., 2007). More recent efforts have extended the initial work on English and addressed other languages such as Chinese (Li et al., 2014), French (Moriceau and Tannier, 2014), Arabic, Italian, Spanish, and Vietnamese (Strötgen et al., 2014a). A study of three domain corpora in English in addition to the newswire domain (SMS, historical narratives and clinical trial abstracts) yielded interesting insight to extend the normalized representation of temporal expressions (Strötgen and Gertz, 2012). This work was then applied to cover historical narratives in an additional seven languages. One key finding was that domain specificity could differ between languages (Strötgen et al., 2014b). This prompts the need to study temporal analysis across domains in a variety of languages in order to adequately characterize each domain and language pairs.

The clinical domain has been addressed during the 2012 i2b2 challenge (Sun et al., 2013b), with a task on temporal relation extraction from clinical narratives. This task used a corpus of clinical notes in English annotated with temporal information (Sun et al., 2013a) based on ISO-TimeML (Pustejovsky et al., 2010). It prompted further work in this domain in English (Jindal and Roth, 2013) and Swedish (Velupillai, 2014), including the release of detailed guidelines for creating temporal annotations of clinical text and a discussion of the clinical domain specificity related to temporal aspects (Styler IV et al., 2014). Finally, clinical TempEval 2015 brought the temporal information extraction tasks of past TempEval campaigns to the clinical domain (Bethard et al., 2015).

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 492–498,
Lisbon, Portugal, 17-21 September 2015.

In this paper, we continue to explore temporal expression identification across domains, with a focus on French narratives. We introduce a new corpus of French clinical narratives annotated with normalized time expressions. We characterize temporal expression recognition in three domains and discuss how the development of an automated temporal expression identification tool may be impacted.

2 Temporal Expression Extraction and Normalization

Rule-based methods were shown to be very efficient for the extraction and normalization of time expressions from news narratives in several languages[1]. In the latest SemEval campaign (UzZaman et al., 2013), the rule-based HeidelTime (Strötgen and Gertz, 2010) out performed machine-learning and hybrid counterparts by a large margin. However, statistical systems obtained promising results with respect to temporal entity extraction.

Based on these results, we chose to use the state-of-the-art rule-based system HeidelTime as well as an in-house statistical tool relying on the Wapiti (Lavergne et al., 2010) implementation of Conditional Random Fields (CRFs) (Lafferty et al., 2001).

Existing HeidelTime settings were used to customize it for the analysis of news and historical narratives in French. In addition, we developed a set of 14 rules to provide additional customization for the analysis of clinical narratives in French.

The CRF model was developed using part of the clinical corpus as a training set, with domain independant surface and lexical features for the text tokens:

- The original token from the text (word form);
- Surface features: capitalization of the token (all in upper/lower case, combination of both), presence of digit (YES, NO) and punctuation mark in the token (PUNCT, NO_PUNCT), temporal type of token according to HeidelTime;
- Lexical features: *n*-grams , number of words, number of digits, number of consecutive repeats. Token frequency was computed based on the entire training corpus.

For clinical text analysis, we experimented with standard tokenization (provided by TreeTagger (Schmid, 1994)) and a custom tokenization where punctuation marks are always considered as token separators, even in dates such as "10-02-2010" or "10.04.10".[2]

3 French Corpora with Temporal Annotations

For this study, we used two available French corpora with TIMEX3 annotations: the French TimeBank corpus (FTB called *news* in this paper) (Bittar et al., 2011) which covers the news domain and the AncientTimes corpus (ATC, called *historical* in this paper) (Strötgen et al., 2014b) which covers the historical domain.

To cover a third domain, we developed a corpus using a set of clinical notes where personal identifying information (PII) had been marked and replaced by surrogates (Grouin and Névéol, 2014). This included marking some temporal expressions such as dates, which were replaced by surrogate dates obtained by substracting a fixed number of days to the original dates. Manual review ensured that there were no format or other errors in the reintroduced surrogate dates. For compatibility with the sources that were already available, including our study corpora and HeidelTime, we chose the TIMEX3 standard for creating temporal annotations.

Three annotators (the authors of this paper) contributed towards the creation of gold-standard annotations for time expressions in the corpus. The annotation of time expressions was carried out in two phases: first, the time expressions and their values were annotated, and second the time expressions were normalized. At the beginning of the first phase, two initial samples of twenty documents were selected for all three annotators to work on. These documents were pre-annotated using the French version of HeidelTime and dates marked as PII. The annotators' task was then to revise the pre-annotations independently. This phase of the annotation process contributed to refining annotation guidelines and creating additional rules to improve on the pre-annotation. Subsequently, the rest of the corpus was divided between annotators, so that each document was annotated independently by two annotators. Final

[1] Normalization is the process of turning any reference to a date into an absolute, formated date.

[2] This tokenization script, adapted from TreeTagger, is available upon request.

gold standard annotations were created by adjudicating any disagreements during meetings between the pair of contributing annotators. Two annotators contributed to the second phase of annotations (normalization). A small sample of 20 documents was annotated independently by the two annotators. Inter-annotator agreement was then computed and found to be sufficiently high to allow for the rest of the corpus to be distributed between annotators.

The phasing of annotations allowed having all corpus documents reviewed several times, so that time expressions that might have been missed during the first phase could be identified and added to the gold standard in the second phase.

To visualize and create annotations, we used the BRAT Rapid Annotation Tool (Stenetorp et al., 2012). Inter-annotator agreement was computed in terms of F-measure, using the companion brateval tool (Verspoor et al., 2013), which we extended to compute inter-annotator agreement on normalized entities.

Table 1 shows the distribution of time expressions according to types in the three corpora used in our study. It suggests that domain specificity is reflected by the types of temporal expressions found in each of the corpora. While *Dates* are prevalent across domains, the news corpus stands out with a high proportion of *Times*, the clinical corpus with a high proportion of *Set* and the historical corpus with almost none of either type. Additional statistics on the clinical corpus are provided in section 4.

	FTB (news)		**ATC (hist.)**		**Clinical**	
	#	%	#	%	#	%
Date	227	53.41	124	81.05	2594	65.14
Dur.	52	12.24	25	16.34	343	8.61
Set	16	3.76	3	0.02	994	24.96
Time	130	30.59	1	0.01	51	1.28

Table 1: Distribution of Time Expressions in three French corpora

4 Results

4.1 French Clinical Corpus with Temporal Annotations

Figure 1 shows an excerpt of the training corpus annotated with temporal expressions. The blue boxes show the normalized value associated with each temporal expression. Due to the confidential nature of the corpus, we are currenlty not able to release it.

Table 2 presents detailed statistics on the clinical corpus. Inter-annotator agreement was .91 F-measure for temporal entity annotation (averaged over the three annotator pairs on the training corpus) and .99 F-measure for temporal normalization (computed on a sample of 20 documents from the training corpus).

	Training	**Test**	**All**
Documents	246	115	361
Tokens	97,008	44,803	141,811
DATE	1,659	935	2,594
DURATION	255	88	343
TEMPORAL SET (Frequency)	605	389	994
TIME	19	32	51

Table 2: Description of the gold standard clinical corpus

4.2 Extraction of Temporal Expressions across Domains

Table 3 presents the results of temporal expression extraction in French narratives across the three domains in our study. The model configurations used are either HeidelTime (H) or statistical (S), adapted to one of the three domains. For HeidelTime models, the adaptation consisted in selecting a domain specific set of rules. We also report results by Strötgen et al. (2014b) showing the difference between HeidelTime 1.5 and the improvements obtained by their new rules for historical French texts. For statistical models, the adaptation consisted in training the model on a corpus of the relevant domain. We studied the effect of corpus size by training a model using a portion of the clinical training data equivalent in size to that of FTB (marked *clin-* in Table 3). However, the ATC corpus was too small to train any usable models (results not shown). Experiments with our adapted, in-house tokenization tool are marked with a + in the models.

The results of the evaluation are reported in terms of precision, recall and F1-measure. We evaluate the extraction of temporal expressions associated to the correct TIMEX3 attribute type (DATE, DURATION, TIME, SET), with the 'strict' measure (only exact match is correct) and the 'relaxed' measures (overlaps are allowed).

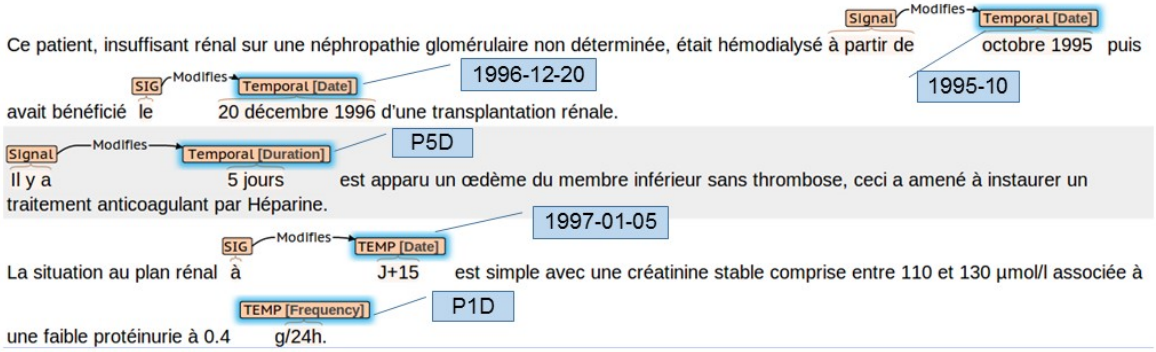

Figure 1: Excerpt from a sample document annotated with temporal expressions; dates and personal health information were replaced by plausible surrogates.

5 Discussion

Overall, our results show that while good performance can be achieved for the extraction of temporal expression on many specific domains, the task of automatically extracting temporal expressions is not solved across the board. Methods that are successfully developed for one specific domain do not carry very well over to other domains without any adaptation work. In our experiments, rule-based methods seem to fare somewhat better in terms of generalizability, but statistical methods can be better optimized for a particular domain, given enough training data. Similar insight resulted from a recent study of negation detection in the clinical domain (Wu et al., 2014). One additional issue highlighted in the negation study was that the definition of the entities that could fall under the scope of negation varied from domain to domain in the available negation corpora. This pitfall is avoided with the temporal task thanks to the use of the TimeML standard. The challenges of extracting temporal expressions across domains that we identified in this study on French correlate well with those described by (Strötgen and Gertz, 2012) on English. The performance of statistical models trained on in-domain data was significantly higher compared to out-domain data: *S-news+* yielded the highest performance on the News corpus (strict F-measure 0.74), and *S-clin+* yielded the highest performance on the Clinical corpus (strict F-measure 0.94).

Adaptation burden. The amount of effort to adapt to new domains was overall limited for the rule-based system: only a few rules needed to be added to the news-oriented HeidelTime (2 for historical, 14 for clinical) to reach comparable performance on other domains. The adaptation effort for the statistical model relies mostly on the availability of annotated corpora in the relevant domains. A competitive CRF model can be trained without using domain specific features. However, we find that adaptation effort can cover pre-processing: for clinical documents, using domain-specific tokenization yielded improved temporal expression extraction for both rule-based and statistical systems. This is in line with previous findings that pre-processing is essential for making experiments reproducible, and variations in pre-processing methods can result in significant difference in performance for different NLP tasks (Fokkens et al., 2013).

Finally, we can note that the difference between HeidelTime and CRFs is much bigger on the strict measure than on the relaxed measure, which may suggest that small variations could be better handled with more covering rules.

Impact of pre-processing, training corpus type and size. Tokenization had a real impact on the performance of statistical models. Using a baseline tokenization method can reduce the performance by several points in F-measure (strict F-measure of 0.68 for *S-news* vs. 0.74 for *S-news+* on the News corpus, strict F-measure of 0.88 for *S-clin* vs. 0.94 for *S-clin+* on the Clinical corpus – see Table 3). Corpus size had a comparably smaller impact. The training dataset used for *S-clin* is 10 times larger than that used for *S-clin-*, and yet it provides an improvement of about 4 points in F-measure for in-domain application. Conversely, for out-domain application, using the larger model is detrimental. Using in-domain training data (or more generally, training

Test	Model	relaxed			strict		
		P	R	F	P	R	F
Clinical	H-news	0.83	0.69	0.75	0.63	0.53	0.57
	H-clin	0.92	0.89	0.90	0.81	0.78	0.79
	S-news	0.86	0.72	0.79	0.60	0.49	0.54
	S-news+	0.86	0.70	0.77	0.64	0.51	0.57
	S-clin-	0.98	0.86	0.92	0.89	0.78	0.83
	S-clin-+	0.98	0.87	0.92	0.95	0.84	0.89
	S-clin	0.98	0.93	0.96	0.91	0.85	0.88
	S-clin+	**0.99**	**0.94**	**0.96**	**0.97**	**0.91**	**0.94**
Historical	H-news1.5*	0.97	0.43	0.59	0.71	0.31	0.43
	H-news*	0.98	**0.84**	**0.91**	**0.89**	**0.76**	**0.82**
	H-clin	0.93	0.44	0.59	0.61	0.40	0.40
	S-news	0.87	0.29	0.44	0.62	0.21	0.31
	S-news+	0.87	0.30	0.45	0.67	0.24	0.35
	S-clin-	0.94	0.28	0.43	0.56	0.17	0.26
	S-clin-+	0.94	0.25	0.40	0.56	0.15	0.24
	S-clin	0.96	0.30	0.46	0.61	0.19	0.29
	S-clin+	**0.99**	0.28	0.43	0.67	0.19	0.29
News	H-news	0.85	**0.79**	**0.82**	**0.83**	**0.78**	**0.81**
	H-clin	0.85	0.79	0.82	0,75	0.70	0.72
	S-news	0.83	0.66	0.68	0.77	0.61	0.68
	S-news+	**0.86**	0.69	0.77	0.83	0.68	0.74
	S-clin-	0.84	0.41	0.55	0.61	0.31	0.41
	S-clin-+	0.76	0.33	0.45	0.55	0.24	0.33
	S-clin	0.76	0.41	0.53	0.62	0.34	0.44
	S-clin+	0.77	0.38	0.51	0.65	0.33	0.43

Table 3: Evaluation of temporal expression extraction in French narratives across three domains. Values from models with a '*' come from Strötgen et al. (2014b). Models with a '+' used custom tokenization. Models with a '-' used the reduced training set.

data that is as close as possible to in-domain, such as News vs. Clinical for Historical) of any reasonable size yields better performance, even if still under the rule-based approach.

Limitations of this study. Size imbalance in the corpora used in our study was a limitation; the clinical corpus is much larger than the other two and the Historical corpus is really small, which limits the applicabiblity of statistical methods. In our work with the clinical corpus, more time was spent on annotating data and implementing statistical models vs. developping rules. Arguably, devoting additional efforts to rule development might improve the rule-based performance.

6 Conclusion and Future Work

This study contributes to a better understanding of temporal expression recognition accross domains. We found that an important part of domain specificity lies in the distribution of the types of temporal expressions accross domains. We also noticed that specific mentions of temporal expressions can be categorized as different types from one domain to another (e.g. *le soir* was generally considered a set in our clinical corpus – as in *every night* – and a time in the news corpus – as in *in the evening*). The results of our domain adaptation experiments suggest that the performance of temporal expression recognition is improved when domain specificity is taken into account by using in-domain training data or domain-specific rules.

In terms of adaptation strategy, our experiments show that the addition of a limited number of rules to the default (news-oriented) HeidelTime leads to matching the expected performance of HeidelTime on a new domain corpus. Furthermore, we show that more substantial efforts spent on annotating data can result in training data that will support a statistical model that outperforms simple rule adaptation. We can hypothesize that devoting equivalent efforts towards rule development may also result in increased performance. We believe that some amount of corpus annotation is necessary to gain adequate corpus knowledge to craft such rules.

Overall, we show that domain adaptation for the extraction of temporal expressions can be done with limited efforts, provided that an adequate corpus is available. We found that the tokenization method used in pre-processing was instrumental for improving statistical model performance across domains.

In future work, we will address the task of temporal expression normalization.

Acknowledgments

This work was supported by the French National Agency for Research under grant CABeRneT[3] ANR-13-JS02-0009-01

The authors thank the Biomedical Informatics Department at the Rouen University Hospital for providing access to the LERUDI corpus for this work.

[3]CABeRneT: Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle

References

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, USA, jun. Association for Computational Linguistics.

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French TimeBank: An ISO-TimeML Annotated Reference Corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 130–134, Stroudsburg, PA, USA. Association for Computational Linguistics.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. In *ACL (1)*, pages 1691–1701, Sofia, Bulgaria. The Association for Computer Linguistics.

Cyril Grouin and Aurélie Névéol. 2014. De-Identification of Clinical Notes in French: towards a Protocol for Reference Corpus Developpement. In *J Biomed Inform*, Aug.

P. Jindal and D. Roth. 2013. Extraction of Events and Temporal Expressions from Clinical Narratives. *Journal of Biomedical Informatics (JBI)*, 10.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Thomas Lavergne, Olivier Cappé, and Fran¸cois Yvon. 2010. Practical Very Large Scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513, Uppsala, Sweden.

Hui Li, Jannik Strötgen, Julian Zell, and Michael Gertz. 2014. Chinese Temporal Tagging with HeidelTime. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 133–137. Association for Computational Linguistics, April.

Véronique Moriceau and Xavier Tannier. 2014. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavík, Iceland, may.

J Pustejovsky, P Hanks, R Saur, A See, R Gaizauskas, A Setzer, D Radev, B Sundheim, D Day, L Ferro, and M Lazo. 2003. The TimeBank corpus. *Corpus Linguistics*, page 647–656.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, (LREC'10)*, pages 394–7, La Valette, Malta, May.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, September.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International op on Semantic Evaluation*, SemEval '10, pages 321–324, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2012. Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Jannik Strötgen, Ayser Armiti, Tran Van Canh, Julian Zell, and Michael Gertz. 2014a. Time for More Languages: Temporal Tagging of Arabic, Italian, Spanish, and Vietnamese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1):1–21.

Jannik Strötgen, Thomas Bögel, Julian Zell, Ayser Armiti, Tran Van Canh, and Michael Gertz. 2014b. Extending HeidelTime for Temporal Expressions Referring to Historic Dates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2390–2397. European Language Resources Association (ELRA), May.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C 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*, page 143–154, Apr.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013a. Annotating temporal information in clinical narratives. *J Biomed Inform*, 46:Suppl:S5–12, Dec.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013b. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc.*, 20:806–813, Sep-Oct.

Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, JF Allen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9. ACL, Jun.

Sumithra Velupillai. 2014. Temporal Expressions in Swedish Medical Text – A Pilot Study. In *Proceedings of BioNLP 2014*, pages 88–92, Baltimore, Maryland, June. Association for Computational Linguistics.

M Verhagen, R Gaizauskas, F Schilder, M Hepple, and J Pustejovsky. 2007. Semeval-2007 task 15: TempEval temporal relation identification. In *SemEval-2007: 4th International Workshop on Semantic Evaluations*.

Karin Verspoor, Antonio Jimeno Yepes, Lawrence Cavedon, Tara McIntosh, Asha Herten-Crabb, Zoë Thomas, and John-Paul Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database (Oxford)*, page bat019–bat019.

S Wu, T Miller, J Masanz, M Coarr, S Halgrim, D Carrell, and C Clark. 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PLoS One*, 9:e112774, Nov.