Detecting Level of Belief in Chinese and Spanish

Juan Pablo Colomer
Columbia University
j.p.colomer@columbia.edu

Keyu Lai
Columbia University
kl2844@columbia.edu

Owen Rambow
CCLS
Columbia University
rambow@ccls.columbia.edu

Abstract

There has been extensive work on detecting the level of committed belief (also known as "factuality") that an author is expressing towards the propositions in his or her utterances. Previous work on English has revealed that this can be done as a word tagging task. In this paper, we investigate the same task for Chinese and Spanish, two very different languages from English and from each other.

1 Introduction: Committed Belief

The term "committed belief" (Diab et al., 2009) has been used to refer to the commitment of a writer towards the propositions she communicates: does she fully believe the proposition, does she believe the proposition may be true, is she reporting someone else's belief without commenting on it, or is she reporting something other than a belief, namely a hope or desire? The notion is closely related to "factuality", which Saurí and Pustejovsky (2009) define as the communicative intention of the writer to make the reader believe what her beliefs are. For a fuller discussion of the relation between the two notions and related notions such as factivity and modality, see (Prabhakaran et al., 2015).

Determining the writer's degree of commitment to the propositions in her text is crucial in understanding text, since if an NLP system fails to identify a proposition as merely wished as opposed to asserted, then clearly the NLP system is failing to understand what is being communicated.

While work on English has been available (both for Committed Belief and for Factuality), no resources have been available for other languages. Recently, the Linguistic Data Consortium (LDC) has annotated small corpora for Chinese and Spanish. This paper summarizes initial systems trained on these corpora.

2 Data

The LDC has released one data set each for Chinese and Spanish committed belief word tagging to the research groups participating in the DARPA DEFT program.[1] The LDC will make this data available to the general research community. We describe these data sets in this section.

2.1 Annotation Scheme

The annotation is a word-based annotation. The goal of the annotation is to identify propositions in the text and to tag them with the degree of Committed Belief. This degree is tagged on the word which is the syntactic head of the proposition. For such syntactic heads, 4 tags are available, they are summarized in Table 1. All words which are not the syntactic heads of propositions get a default "O" tag (or "Other"). We only evaluate our performance on the four belief tags.

This annotation scheme extends the annotation scheme proposed by Diab et al. (2009) by splitting its NCB tag into NCB and ROB. (In the scheme of (Diab et al., 2009), our NCB and ROB were combined because in both tags we cannot infer a committed belief of the writer; however, in terms of our knowledge of the writer's cognitive state, they clearly represent very different categories.) For a fuller discussion of the tagsets, see (Werner et al., 2015).

[1] LDC2015E99 for Chinese and LDC2016E40 for Spanish.

Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics,
pages 22–30, Osaka, Japan, December 12 2016.

Tag	Meaning	Example
CB	Committed Belief	John will arrive tomorrow
NCB	Non-committed Belief	John may arrive tomorrow
ROB	Reported Belief	Mary says that John will arrive tomorrow
NA	Not a Belief	I hope that John will arrive tomorrow

Table 1: Explanation of the four belief tags used in the annotation, along with English examples

2.2 Chinese Data

The Chinese corpus is sampled from Chinese Discussion forums. The topics mostly focus on politics and news stories. The corpus is annotated at the character level, not the word level. To annotate what would be considered a word, the corpus uses the label of the first annotated character as the label for the whole word. For example, if the word 访问'access' is the head of a proposition in which the author expresses committed belief, then the annotation is "访/CB问/O" rather than "访问/CB". The character 访is annotated as CB rather than the word 访问because the Committed Belief annotation did not want to have to perform word segmentation as part of the annotation task, which can be a time consuming (and not always obvious) task. As a result, the annotation scheme is compatible with different choices as to word segmentation.

We do perform word segmentation in this work, using the Stanford tools (Manning et al., 2014). When we do word segmentation, and if at least one character has an annotation, then that annotation is carried over to the whole word. If all characters comprised by the word don't have annotations, then the word remains unlabelled (i.e., it gets the O tag). It did not happen in our corpus that more than one character in the same word received tags which were contradictory. We compare using characters and using words in Section 4.1.

We divided the whole corpus into 80% training set, 10% development set and 10% test set in term of characters for further experiment. The numbers of characters in each subsets are: training set: 96735; development set: 11747; test set: 12155. Here is a simple example:

(1) 妈妈 说/CB 我 看/ROB 他 喜欢/ROB 吃 这个
Mom say I think he like eat this
'Mom said I think he likes eating this'

	Training	Development	Test	Training	Development	Test
	Chinese			Spanish		
CB	7,939 (13%)	974 (14%)	1,076 (15%)	4,563 (7%)	496 (7%)	600 (7%)
NA	5,294 (9%)	639 (9%)	492 (7%)	3,288 (5%)	406 (6%)	494 (6%)
NCB	209 (0%)	39 (1%)	24 (0%)	267 (0%)	32 (0%)	46 (1%)
ROB	1,086 (0%)	72 (1%)	57 (1%)	437 (1%)	13 (0%)	47 (1%)
O	44,406 (75%)	5,432 (76%)	5,621 (77%)	55,215 (87%)	6,228 (86%)	7,448 (86%)
Total	58,934	7,156	7,270	63,770	7,175	8,635

Table 2: Chinese and Spanish words per label and data set

2.3 Spanish Data

The Spanish corpus was also extracted from discussion forums. It is important to note that people from different Spanish speaking countries write in these forums. Also, they tend to use an informal language, thus there is a significant diversity of slang words which makes the task hard even for a Spanish native speaker.

The corpus was separated approximately into 80% training set, 10% development set and 10% test set in terms of labeled words.

Example:

(2) Creo/CB que debería/NCB haberlo escrito/CB en mayúscula
think that should have-it written in uppercase
'I think I should have written it in capital letters'

2.4 Discussion of Data Sets

The data sets are summarized in Table 2. Several observations are in order:

- For each languages, the distribution of the labels is fairly similar in the training, development, and test sets.
- In Spanish, fewer words are tagged with belief labels (i.e., more words are tagged with O). This is because Spanish has determiners, auxiliaries, and in general more function words which do not receive Committed Belief labels.
- In both languages, there are very few NCB and ROB tags (with ROB more frequent than NCB). As we will see, these tags are accordingly hard to predict.

Since the information about Committed Belief is expressed as tags on words, we can define the task as a word-tagging task, as was also done previously for English.

3 Features and Experimental Setup

3.1 Features Used in Both Languages

In our analysis we used some common features for both languages. These features are the following:

- Word: One-hot encoding representation.
- Part-of-Speech (POS): One-hot encoding representation (using different tagsets for the two languages of course). The use of POS is motivated by the need to find the syntactic heads of propositions, which are typically verbs.
- 64 dimension word embedding: We used Polyglot (Al-Rfou et al., 2013) to get the word embedding for the two languages.

For word segmentation in Chinese and POS-tagging in Chinese and Spanish, we used the Stanford tools (Manning et al., 2014).

Additionally, the process to obtain the features vector of a word is the same on Chinese and Spanish. We experimented with 3 configurations of context windows to compute above features; they differ in where in the 5-word context window the target word is found. Let w_0 be the target word and w_i the word in i-th position relative to w_0.

- [-2/+2]: $[w_{-2}, w_{-1}, w_0, w_1, w_2]$
- [-3/+1]: $[w_{-3}, w_{-2}, w_{-1}, w_0, w_1]$
- [-4/0]: $[w_{-4}, w_{-3}, w_{-2}, w_{-1}, w_0]$

Thus, the feature vector of the target word is formed by stacking the features' representations of all words in the context window.

3.2 Features Used only in Spanish

In addition to the features described in 3.1 an important feature for Spanish is the lemma of a word. The software used to extract these features is Freeling 3.0 (Padró and Stanilovsky, 2012).

3.3 Baseline

We will consider our baseline to be a system that trains only on words and uses the [-2,+2] context window (i.e., the words are chosen to be centered on the context word). We also consider this our baseline for Chinese, even though it requires the additional step of word segmentation. This is because a character-based model performs much worse, as we will see in Sectionsec:ch-ch-w.

3.4 Experimental Setup

For both languages, we trained and fine-tuned the parameters of a Linear SVM classifiers from Scikit-learn library (Pedregosa et al., 2011). This classifier implements a one-vs-all strategy which has a similar performance as an SVM classifier with one-vs-one strategy, but its runtime is considerably faster.

We report results only on the tags for the heads of propositions (i.e., not on the O tag). We use F-measure to report results, and used a weighted average F-measure to summarize the results.

4 Chinese Results

4.1 Characters or Words?

Chinese is typically written without spaces between two words (different from European languages including Spanish and English). The identification of words in Chinese is a typical initial processing step in Chinese NLP. However, since the annotation is in fact at the character level (see Section 2.2), we perform experiments to see if annotation at the character level performs better than at the word levels. We use two windows for the character experiments, namely [-2,+2] (a five-character window centered on the target character) and [-4,+4] (a nine-character window centered on the target character). For the word experiments, we use the baseline configuration (only words, with a [-2,+2] context window.

The results are shown in Table 3. As can be seen, using words far outperforms characters, even if we use a much larger window for characters than for words. We therefore use words for the remainder of our experiments.

	Characters [-2/+2]			Characters [-4/+4]			Words [-2/+2]		
	prec.	recall	f1	prec.	recall	f1	prec.	recall	f1
CB	0.4008	0.4714	0.4333	0.4406	0.4629	0.4515	0.5533	0.5329	**0.5429**
NA	0.3539	0.4059	0.3781	0.3798	0.4088	0.3937	0.5129	0.4351	**0.4708**
NCB	0.0353	0.1429	0.0566	0.0567	0.1905	0.0874	0.0968	0.2308	**0.1364**
ROB	0.0349	0.1579	0.0571	0.0353	0.1579	0.0577	0.0723	0.2361	**0.1107**
Wted Avg.	0.3601	0.4266	0.3888	0.3926	0.4240	0.4056	0.5083	0.4772	**0.4891**

Table 3: Chinese: Comparison of labeling characters with labeling words (after word segmentation); first six result columns are based on characters with different context windows, next three columns are based on words. Boldface indicates the best F1-measure performance per label across the three experiments.

4.2 Adding POS

We now investigate the role of part-of-speech (POS) tags. We first use the same window as in the baseline (the last experiment in Table 3), namely [-2,+2], i.e., the target word and two words to the left and two words to the right. The results for the same window are shown in the first three result columns in Table 4. Comparing to the word results from Table3, we see that the addition of POS increases the results for the common tags CB and NA as well as for ROB by around 2% absolute; NCB is not affected. We therefore keep POS tags in all subsequent experiments.

In a second round of experiments we vary the context window. In the results for [-3,+1], we let the target word be the third word in the 5-word window (middle three result columns in Table 4), and then we consider the [-4,0] window in which the target word is the last word in the 5-word window (last three result columns in Table 4). We see that except for ROB, the best performance is always obtained using a window centered on the target word.

4.3 Using Word Embeddings

Finally, we add word embeddings to the word and POS features. We again experiment with the position of the target word in the context window. The results are shown in Table 5. We see that when we use word embeddings, the left context becomes more valuable than the right context, and we now obtain better results if we use context window [-3,+1] (i.e., the target word is in position 4 of the 5-word

	Words and POS [-2/+2]			Words and POS [-3/+1]			Words and POS [-4/0]		
	prec.	recall	f1	prec.	recall	f1	prec.	recall	f1
CB	0.5643	0.5719	**0.5681**	0.5708	0.5585	0.5646	0.5494	0.5370	0.5431
NA	0.5273	0.4679	**0.4959**	0.5083	0.4789	0.4932	0.4903	0.4726	0.4813
NCB	0.1000	0.2051	**0.1345**	0.0964	0.2051	0.1311	0.0882	0.2308	0.1277
ROB	0.0872	0.2639	0.1310	0.1005	0.3056	**0.1512**	0.0756	0.2500	0.1161
Wted AVG	0.5206	0.5120	**0.5134**	0.5175	0.5103	0.5112	0.4976	0.4942	0.4932

Table 4: Chinese: Using POS tags, experimenting with different positions for the target word w_0 in the window. Boldface indicates the best F1-measure performance per label across the three experiments.

context window). These results are slightly better for all labels compared to the best results without word embeddings; for ROB, they are only slightly worse.

	Words, POS, and Embedding [-2/+2]			Words, POS, and Embedding [-3/+1]		
	prec.	recall	f1	prec.	recall	f1
CB	0.5518	0.5688	0.5602	0.5647	0.5780	**0.5713**
NA	0.5217	0.4898	0.5052	0.5078	0.5086	**0.5082**
NCB	0.0667	0.1282	0.0877	0.1061	0.1795	**0.1333**
ROB	0.0675	0.2222	0.1036	0.0940	0.3056	**0.1438**
Weighted AVG	0.5100	0.5151	0.5104	0.5139	0.5319	**0.5204**

Table 5: Chinese: Using word embeddings, with context window [-2,+2] (word in position 3 of 5-word context window, first three result columns) and context window [-3,+1] (word in position 4 of 5-word context window, last three result columns). Boldface indicates the best F1-measure performance per label across the three experiments.

5 Spanish Results

5.1 Lexical Features

We start out our experiments on the development set by using only lexical features, and we vary the context window. As can be seen from the results in Table 6, the best results for the common labels CB and NA are obtained for context window [-2,+2] (i.e., the target word is centered in the window), while the rarer labels ROB and NCB, performing far worse overall, profit from a greater left context window. The effect is particularly strong for ROB, presumably because the larger left context allows the system to detect verbs of attribution (or perhaps the subordinating conjunction *que* 'that').

	Words [-2/+2]			Words [-3/+1]			Words [-4/0]		
	prec.	recall	f1	prec.	recall	f1	prec.	recall	f1
CB	0.6853	0.5444	**0.6067**	0.6559	0.5343	0.5889	0.6432	0.5343	0.5837
NA	0.6949	0.5665	**0.6242**	0.6787	0.5567	0.6116	0.6817	0.5222	0.5914
NCB	0.3636	0.1250	0.1860	0.3077	0.1250	0.1778	0.4000	0.1250	**0.1905**
ROB	0.0625	0.0769	0.0690	0.0667	0.0769	0.0714	0.1538	0.1538	**0.1538**
Wted AVG	0.6700	0.5333	**0.5926**	0.6458	0.5238	0.5776	0.6448	0.5101	0.5678

Table 6: Spanish: Word Features. Boldface indicates the best F1-measure performance per label across the three experiments.

5.2 Adding POS and Lemmas

In Table 7, we add POS tags as features. We see that results improve across the board. For the common tags CB and NA, the best results continue to be obtained from a the [-2,+2] context window centered on

the target word, while ROB and NCB still profit from more left context.

Lemmas can be a way of reducing data sparseness in highly inflected languages, since they collapse all inflected forms of a lexeme to a single representative. Results using words, POS tags, and lemmas are shown in Table 8. We see only relatively small changes resulting from the use of lemmas. For reasons that are not clear to us, the [-3,+1] context window now performs best on average as well as for the specific tags CB, NCB, and ROB. For the NA tag, even more left context is useful, with the [-4,0] context window performing best. When comparing the best results per label to the best results per label without lemmas (Table 7), we see that the use of lemmas increases the performance for all labels except ROB. However, because the best performance with lemmas is achieved using different configurations, the weighted average does not improve through the use of lemmas.

	Words and POS [-2/+2]			Words and POS [-3/+1]			Words and POS [-4/0]		
	prec.	recall	f1	prec.	recall	f1	prec.	recall	f1
CB	0.6681	0.6411	**0.6543**	0.6542	0.6371	0.6456	0.6475	0.6593	0.6533
NA	0.6987	0.6798	**0.6891**	0.6759	0.6576	0.6667	0.6856	0.6552	0.6700
NCB	0.2857	0.1250	0.1739	0.3571	0.1563	**0.2174**	0.3636	0.1250	0.1860
ROB	0.1176	0.1538	0.1333	0.1176	0.1538	0.1333	0.2222	0.1538	**0.1818**
Wted AVG	0.6607	0.6336	**0.6458**	0.6461	0.623	0.6331	0.6484	0.6325	0.6382

Table 7: Spanish: Using words and POS tags. Boldface indicates the best F1-measure performance per label across the three experiments.

	Words, POS, Lemma [-2/+2]			Words, POS, Lemma [-3/+1]			Words, POS, Lemma [-4/0]		
	prec.	recall	f1	prec.	recall	f1	prec.	recall	f1
CB	0.6762	0.6190	0.6463	0.6716	0.6431	**0.6571**	0.6522	0.6351	0.6435
NA	0.6738	0.6970	0.6852	0.6801	0.6650	0.6725	0.6977	0.6823	**0.6899**
NCB	0.2500	0.1250	0.1667	0.3750	0.1875	**0.2500**	0.3125	0.1563	0.2083
ROB	0.0952	0.1538	0.1176	0.1667	0.1538	**0.1600**	0.1538	0.1538	0.1538
Wted AVG	0.6528	0.6294	0.6395	0.6583	0.6304	**0.6431**	0.6534	0.6326	0.6420

Table 8: Spanish: Adding Lemmas. Boldface indicates the best F1-measure performance per label across the three experiments.

5.3 Using Word Embeddings

We finally add word embeddings (retaining words, POS, and lemmas). The results are shown in Table 9, for the three context windows. We observe that the best window configuration differs even more by label than before, with NA preferring a balanced left and right context.

The best performing single configuration (as measured by weighted average) is [-3,+1], i.e., the target word in the 4th position in the 5-word window, which is also our overall best performing configuration for Spanish.

6 Results on Test Sets

We apply the best performing configurations of each language to the respective held-out test sets, with the results shown in Table 10. We see for both languages a decrease compared to the best result on the development set, of 4% absolute for Chinese, and 7% absolute for Spanish. Presumably this is at least partially due to overfitting to the development set.

7 Discussion

We have trained belief taggers for Chinese and Spanish. Results on the development sets show striking similarities between the two languages:

	Words, POS, Lemmas, Embeddings [-2/+2]			Words, POS, Lemmas, Embeddings [-3/+1]			Words, POS, Lemmas, Embeddings [-4/0]		
	prec.	recall	f1	prec.	recall	f1	prec.	recall	f1
CB	0.6763	0.6149	0.6441	0.6709	0.6452	0.6578	0.6739	0.6472	**0.6598**
NA	0.6872	0.6872	**0.6872**	0.6990	0.6576	0.6777	0.6959	0.6650	0.6801
NCB	0.3529	0.1875	0.2449	0.4	0.25	**0.3077**	0.375	0.1875	0.25
ROB	0.0909	0.1538	0.1143	0.125	0.1538	**0.1379**	0.1053	0.1538	0.125
Wted AVG	0.6620	0.6251	0.6430	0.6663	0.6304	**0.6474**	0.6654	0.6325	0.6473

Table 9: Spanish: Using word embeddings. Boldface indicates the best F1-measure performance per label across the three experiments.

Chinese: Test Results for Features: Word, Part-Of-Speech, Word Embedding on context window [-3,+1]

	precision	recall	f1-score
CB	0.5581	0.5000	0.5275
NA	0.4016	0.5142	0.4510
NCB	0.0118	0.0417	0.0183
ROB	0.0484	0.2105	0.0787
Weighted AVG	0.4858	0.4876	0.4818

Spanish: Test Results for Features: Word, Part-Of-Speech, Lemma, Word Embedding on context window [-3,+1]

	precision	recall	f1-score
CB	0.598	0.605	0.6015
NA	0.6195	0.6559	0.6372
NCB	0.1053	0.0435	0.0615
ROB	0.0833	0.0426	0.0563
Weighted AVG	0.5675	0.5822	0.5738

Table 10: The results of the best configurations on the test sets

- For both languages, the best configuration includes word, POS, and word embeddings, using context window [-3,+1] (in which the target word in the 4th position of the 5-word context window).
- For both languages, the major increase over using only words comes from POS tags. This is plausible since they help the tagger identify the syntactic heads of propositions (which need to be tagged with a belief tag).
- For both languages, word embeddings help a small amount. The relatively small contribution from the word embeddings may be due to the fact that the word embeddings do not capture the right generalizations for this task, or they are trained on corpora that are too small or not representative of our corpora.
- For both languages, the distribution of the tags is fairly similar, with the result that the rare tags NCB and ROB are predicted badly.
- The use of lemmas for Spanish does not contribute much.

There are also some interesting differences between the languages.

- For each tag, the Chinese results are inferior to the Spanish results, except tag ROB. We have no explanation for the fact that ROB performs better in Chinese than in Spanish.
- The differences in performance between Chinese and Spanish are particularly large (in relative terms) for NA and NCB. These are two types of belief which in Spanish are often signaled in the inflections. For example, NAs are often signaled by infinitives which are complements of verbs of obligation (*tiene que transofrmarla*) or wishing (*quiere sostener*), and NCBs are signaled by infinitives after modal verbs (*debe sentir*).

(3) a. tiene/CB que transformarla/NA si quiere/NA sostener/NA el negocio
must transform-it if desires sustain the business
'He must transform it if he wants to sustain business'

b. uno se debe sentir/NCB un verdadero boludo
one oneself must feel a real idiot
'One must feel like a real idiot'

We hypothesize that NA and NCB are specifically helped by the Spanish morphology as captured in the POS tags. This hypothesis is also supported when we consider the error reduction achieved by adding POS to the word feature only. For Chinese, the error reduction is 5.5% for CB and 4.7% for NA (derived from Tables 3 and 4), while for Spanish the error reduction is 12.1% for CB and 17.2% for NA (derived from Tables 6 and 7), suggesting that Spanish profits more from POS tags than Chinese, and crucially, Spanish NA profits more than Spanish CB.

When we compare these results to the English results reported by Prabhakaran et al. (2010), we see that without syntactic features (which we do not use in this paper), their numerical results are somewhat similar to ours. While their training set is much smaller (around 10,000 words, only a sixth of our training corpora), the results using only lexical features and POS tags are similar to ours (57% F-measure weighted average). However, when features derived from a parse tree are derived, the score goes up by 7% absolute. This is because NCB, ROB, and NA labels often correspond to syntactic configurations involving bi-clausal structures, and require an exact analysis of the lexicon-syntactic structure. This is true not only of English, but also of Chinese and Spanish. We intend to incorporate parsing in future work.

8 Future Work

We have seen that we can predict Committed Belief in Chinese and Spanish with acceptable accuracy for the common labels of CB and NA. Future work will concentrate on using a parser, which we expect to boost performance considerably.

Acknowledgments

We would like to thank the four anonymous reviewers for their thoughtful and useful comments; unfortunately, we have not been able to respond to all of the comments in this version of the paper.

Rambow's work on this paper was supported by the DARPA DEFT Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022, Beijing, China, August. Coling 2010 Organizing Committee.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado, June. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268. 10.1007/s10579-009-9089-9.

Gregory Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. 2015. Committed belief tagging on the factbank and lu corpora: A comparative study. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 32–40, Denver, Colorado, June. Association for Computational Linguistics.