A Crowdsourced Frame Disambiguation Corpus with Ambiguity

Anca Dumitrache
FD Mediagroep,
Vrije Universiteit Amsterdam
Netherlands
anca.dmtrch@gmail.com

Lora Aroyo
Google
USA
l.m.aroyo@gmail.com

Chris Welty
Google
USA
cawelty@gmail.com

Abstract

We present a resource for the task of FrameNet semantic frame disambiguation of over 5,000 word-sentence pairs from the Wikipedia corpus. The annotations were collected using a novel crowdsourcing approach with multiple workers per sentence to capture *inter-annotator disagreement*. In contrast to the typical approach of attributing the best single frame to each word, we provide a list of frames with disagreement-based scores that express the confidence with which each frame applies to the word. This is based on the idea that inter-annotator disagreement is at least partly caused by ambiguity that is inherent to the text and frames. We have found many examples where the semantics of individual frames overlap sufficiently to make them acceptable alternatives for interpreting a sentence. We have argued that ignoring this ambiguity creates an overly arbitrary target for training and evaluating natural language processing systems - if humans cannot agree, why would we expect the correct answer from a machine to be any different? To process this data we also utilized an expanded lemma-set provided by the Framester system, which merges FN with WordNet to enhance coverage. Our dataset includes annotations of 1,000 sentence-word pairs whose lemmas are not part of FN. Finally we present metrics for evaluating frame disambiguation systems that account for ambiguity.

1 Introduction

Crowdsourcing has been a popular method to collect corpora for a variety of natural language processing tasks (Snow et al., 2008), although one of its downsides is the crowd's lack of domain knowledge that is helpful in solving some tasks. *Semantic frame disambiguation* is an example of a complex natural language processing task that is usually performed by linguistic experts, subjected to strict annotation guidelines and quality control (Baker, 2012). The theory of frame semantics (J Fillmore, 1982) defines a *frame* as an abstract representation of a word sense, describing a type of entity, relation, or event, together with the associated *roles* implied by the frame. The FrameNet (FN) corpus (Baker et al., 1998) is a collection of semantic frames, together with a corpus of documents annotated with these frames. Similarly to word-sense disambiguation, *frame disambiguation* is the task of obtaining the correct frame for each word, since many words have multiple possible meanings.

Using domain experts for frame disambiguation is expensive and time consuming, resulting in small corpora for this task that do not scale well for modern machine learning methods – FN version 1.7, the latest one at the time of writing, contains only about 10,000 sentences annotated with frames. Furthermore, only using one expert to perform the annotation makes it difficult to capture any diversity of perspectives.

There have been a number of small-scale attempts at using crowdsourcing for frame disambiguation in sentences, showing that the crowd has comparable performance to the FN domain experts (Hong and Baker, 2011), and that the crowd can be used to correct wrong examples that have been collected automatically (Pavlick et al., 2015). Crowd performance can be improved by combining frame role identification with disambiguation (Fossati et al., 2013), or by asking crowd workers to give each other feedback and then letting them change their answer (Chang et al., 2015). Crowdsourcing has also been useful to identify the ambiguity in frame disambiguation (Jurgens, 2013).

Previously, we have shown (Dumitrache et al., 2018a) that while the crowd and FN expert mostly agree over frame disambiguation, disagreement cases are often caused by ambiguity, such as vague or overlapping frame definitions, or incomplete

Proceedings of NAACL-HLT 2019, pages 2164–2170
Minneapolis, Minnesota, June 2 - June 7, 2019.

information in the sentence. Because of these issues with the input data, the approach of selecting one single correct frame for every word, and ignoring alternative interpretations, often results in arbitrary, incomplete ground truth corpora. In order to aggregate annotated data while preserving disagreement, we use the CrowdTruth method[1] (Aroyo and Welty, 2014), which encourages using multiple crowd annotators to perform the same work, and processes the disagreement between them to signal low quality workers, sentences, and frames.

This paper presents a crowdsourced FN frame disambiguation corpus of 5,042 sentence-word pairs (which has since grown to over 9,000 since the submission of this paper). More than 1,000 of these are lexical units (LUs) not part of FN. To our knowledge, it is the largest corpus of this type outside of FN. In addition, we applied the CrowdTruth method, in which each sentence and lexical item is accompanied by *a list of multiple frames* with scores that express the confidence with which each frame applies to the word. This allows us to demonstrate that ambiguity is a prominent feature of frame disambiguation, with many cases where more than one possible frame can apply to the same word. Finally, we present an evaluation of several frame disambiguation models using evaluation metrics that leverage the multiple answers and their confidence scores, and show that even a model that always predicts the top crowd answer will not always have the best performance.

2 Corpus Collection & Analysis

2.1 Data Preprocessing

Our corpus consists of 5,042 candidate word-sentence pairs from Wikipedia (which has since grown to over 9,000 since the submission of this paper) and a candidate list of frames for the word, with 742 unique frames and 1,705 unique lexical units (LUs). The sentences have been randomly selected, based on these criteria:

- The candidate word has *no more than 25 candidate frames*, to not overwhelm the annotators.
- The part of speech of the word is a *verb*.
- The *distribution of candidate frames was optimized for maximum diversity* using a greedy approach.

To gather the candidate frames for each word, we gathered the candidate frames associated with the LU from FN1.7. Next we completed the candidate list using Framester (Gangemi et al., 2016), which maps FN semantic frames to synonym sets from WordNet (Miller, 1995). The sentences were processed with tokenization, sentence splitting, lemmatization and part-of-speech tagging. Then each word with a frame attached to it was matched with all of its possible synonym sets from WordNet, while making sure that the part-of-speech constraint of the synonym set is fulfilled. Using the WordNet mapping, we constructed the list of additional candidate frames for each word. Framester disambiguation used release 1.5 of FN, and some frames changed names in version 1.7, so we manually mapped these frames from FS to their latest version. Framester disambiguation was also used to collect a subset of our corpus consisting of 1,000 sentence-word pairs with LUs that are not part of the FN corpus. For simplicity, we refer to the sentence-word pairs as sentences in the rest of the paper.

2.2 Crowdsourcing Setup

We ran the task on Amazon Mechanical Turk, where the workers were asked to select *all frames* that fit the sense of the highlighted word in a sentence from the multiple choice candidate list, or that none of the frames is correct. We used 15 workers/sentence that were paid $0.05 for each judgment, and a total cost of $1.35 per sentence (after factoring in the additional AMT costs).[2]

To aggregate the results of the crowd while also capturing inter-annotator disagreement, we use the CrowdTruth metrics[3] (Dumitrache et al., 2018b), replicating the setup from our previous work (Dumitrache et al., 2018a). The choice of frames of one worker over one sentence are aggregated into a *worker vector* – a binary vector with $n + 1$ components, where n is the number of frames shown together with the sentence, where the decision to pick each of the frames (or none) corresponds to a component in the vector. The vectors are used to calculate quality scores for workers, sentences

[1] http://crowdtruth.org

[2] https://mturk.com/

[3] https://github.com/CrowdTruth/CrowdTruth-core

#	SENTENCE	SQS	FRAMES (FSS)
1	Domestication of plants has, over the centuries **improved** disease resistance.	0.652	*improvement or decline* (0.823), *cause to make progress* (0.683)
2	He is the 5th of 8 male players in history to **achieve** this.	0.626	*accomplishment* (0.764), *successful action* (0.709)
3	Albertus Magnus, a Dominican monk, **commented** on the operations and theories of alchemical authorities.	0.511	*communication* (0.522), *statement* (0.703)
4	He **slices** at Hector's armor, throwing him off guard and spinning him around.	0.319	*part piece* (0.499), *cause harm* (0.4), *cutting* (0.394), *attack* (0.254), *hit target* (0.227)
5	Another 46 steps **remain** to climb in order to reach the top, the "terrasse", from where one can enjoy a panoramic view of Paris.	0.308	*left to do* (0.497), *remainder* (0.478), *state continue* (0.319), *existence* (0.155)
6	Borzoi males frequently **weigh** more.	0.283	*assessing* (0.421), *dimension* (0.402), *importance* (0.128)
7	The dance includes bending and **straightening** of the knee giving it a touch of Cuban motion.	0.24	*reshaping* (0.495), *arranging* (0.356), *body movement* (0.298), *cause motion* (0.249)

Table 1: Example sentences with disagreement over the frame annotations (candidate word in bold).

and frames. Although we make all quality scores available as part of the corpus, in this paper we focus on:

- **frame-sentence score (FSS):** the degree with which a frame matches the sense of the word in the sentence. It is the ratio of workers that picked the frame to all the workers that read the sentence, weighted by the worker quality. A high FSS means the frame is clearly expressed in a sentence.

- **sentence quality (SQS):** the overall worker agreement over one sentence. It is the average cosine similarity over all worker vectors for one sentence, weighted by the worker quality and frame quality. A high SQS indicates a clear sentence.

The aggregated crowdsourcing results and the FN 1.5 to 1.7 mapping table are available online.[4]

2.3 Ambiguity in the Corpus

An analysis of the corpus found many examples of inter-annotator disagreement, of which a few examples are shown in Table 1. For 720 sentences, a majority of the workers picked at least 2 frames (examples 1-3 in Table 1). And for 1,514 sentences, no one frame has been picked by a majority of the workers (examples 4-7 in Table 1). Disagreement is also more prominent in the sentences where the LU is not a part of FN (Figure 1).

The disagreement comes from a variety of causes: a parent-child relation between the frames (*statement* and *communication* in #3), an overlap in the definition of the frames (*accomplishment* and *successful action* in #2), the meaning of the word is expressed by a composition of frames (in #7, "straightening of the knee" is a combination of *reshaping* the form of the knee, *arranging* the knee in the right position, and *body movement*), and combinations of all of these reasons (in #4, "slices" is a combination of *part piece* and *cause harm*, and the other frames are their children). More example sentences for each type of disagreement are available in the appendix. The sentences themselves are not difficult to understand, and it can be argued that all of them have one frame that applies the best for the word. The goal of this corpus is to show that next to this best frame for the word, there are other frames that apply to a lesser degree, or capture a different part of the meaning. When evaluating a model for frame disambiguation, it seems unfair to penalize misclassifications of frames that still apply to the word, but with less clarity, in the same way we would penalize a frame that captures a wrong meaning. Also, we argue that models should take into account that annotators do not agree over some examples, and treat them differently than clear expressions of frames. Disagreement can also be caused by worker mistakes (in #6, *dimension* refers to the size of the object, not the act of measuring the size). While we try to mitigate for this by weighing confidence scores with the worker quality, the mistakes still appear in the corpus. This type of disagreement could be useful in future work to identify examples that workers need to be trained on.

3 Evaluating Frame Disambiguation

3.1 Systems Tested

As an example usage of our corpus, we used it to evaluate these frame disambiguation models:

[4] https://github.com/CrowdTruth/FrameDisambiguation

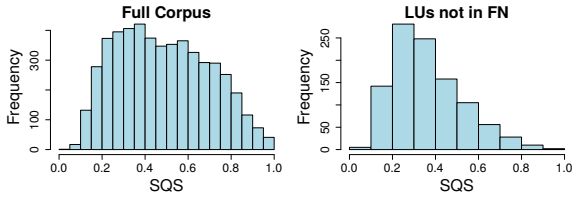

Figure 1: Histogram of SQS values - the quality scores in sentences where the LU is not in FN skew lower.

1. **OS:** The Open-Sesame (Swayamdipta et al., 2017) classifier, pre-trained on the FN corpus (release 1.7). Given a word-sentence pair, OS uses a BiLSTM model with a softmax final layer to predict a single frame for the word. If the LU is not in FN, it cannot make a prediction.

2. **OS+:** We modified the OS classifier to perform multi-label classification. To calculate the confidence score for candidate frame f, we removed the softmax layer and passed the output of the BiLSTM model $\nu(f)$ through the following transformation: $c(f) = [1 + tanh\ \nu(f)]/2$. This gave a score $c(f) \in [0, 1]$ expressing the confidence that frame f is expressed in the sentence.

3. **FS:** Framester includes a tool for rule-based multi-class multi-label frame disambiguation (Gangemi et al., 2016). While for the dataset pre-processing (Sec. 2) we considered the frames for all synsets a word is part of, FS performs an additional word-sense disambiguation step to return a more precise list of frames. We used the tool with *profile T* as it was shown to have the overall better performance. FS can only predict FN frames from the 1.5 release, which is missing 202 frames from version 1.7.

While OS+ produces confidence scores, the other methods produce binary labels for each frame-sentence pair. These models do not have state-of-the-art performance (Hermann et al., 2014; FitzGerald et al., 2015), we picked them because they were accessible and allowed testing on a novel corpus. Finally, we evaluate the quality of the **TC** corpus, containing only the top frame picked by the crowd for every sentence. This test shows what is the best possible performance over our corpus that can be expected from a system such as OS that selects a single frame per sentence.

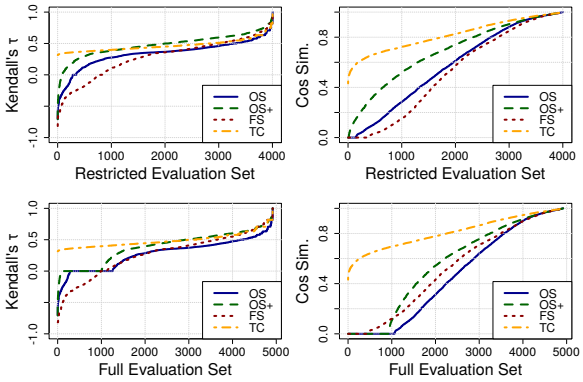

Figure 2: Baselines evaluation results.

	Eval. Metric	OS	OS+	FS	TC
R-Set	Kendall's τ AUC	0.339	**0.477**	0.279	0.466
	Kendall's τ w-avg	0.362	**0.497**	0.3	0.48
	Cos Sim AUC	0.57	**0.685**	0.518	0.818
	Cos Sim w-avg	0.608	**0.717**	0.545	0.854
F-Set	Kendall's τ AUC	0.269	**0.379**	0.253	0.491
	Kendall's τ w-avg	0.307	**0.421**	0.284	0.501
	Cos Sim AUC	0.453	**0.544**	0.511	0.810
	Cos Sim w-avg	0.515	**0.607**	0.539	0.849

Table 2: Aggregated evaluation results.

3.2 Evaluation Metrics & Results

Instead of traditional evaluation metrics that require binary labels, we propose an evaluation methodology that is able to consider multiple candidate frames for each sentence and their quality scores. We use **Kendall's** τ list ranking coefficient (Kendall, 1938) and **cosine similarity** to calculate the distance between the list of frames produced by the crowd labeled with the FSS, and the frames predicted by the baselines in each sentence. Whereas Kendall's τ only accounts for the ranking of the FSS for each frame, cosine similarity uses the actual FSS values in the calculation of the similarity. Both metrics compute a score per sentence (Kendall's $\tau \in [-1, 1]$, and cosine similarity $\in [0, 1]$). Using these metrics, we produce two aggregate statistics over our test corpus: (1) the area-under-curve (AUC) for each metric, normalized by the corpus size, and (2) the SQS-weighted average of each metric ($w-avg$), which also accounts for the ambiguity of the sentence as expressed by the SQS. We evaluate on two versions of the corpus: (1) the restricted set (R-Set) of 4,000 sentences with LUs from the FN corpus, and (2) the full set (F-Set) of 5,042 sentences.

The results (Figure 2 & Table 2) show that OS+ performs best out of all the models, even taking into account sentences with LUs not in FN for which OS+ cannot disambiguate. FS performs the

worst out of all models on R-SET, because it cannot find newly added frames from the latest FN release, but improves on the F-SET (FS can find candidate frames for LUs not in FN). The scores on the F-SET were lower for all baselines, suggesting that sentences with LUs not in FN are more difficult to classify – this could be because FN is missing frames that can express the full meaning of these LUs. TC has a good performance, but is far from being unbeatable – when measuring Kendall's τ over the R-SET, OS+ performs better than TC.

4 Conclusions

We described a FrameNet frame disambiguation resource of 5,042 sentence-word pairs, and 1,000 LUs that are new to FN – the largest corpus of this type outside of FN. Since the submission of this paper, the corpus has grown to over 9,000 sentence-word pairs. We also provide confidence scores for each candidate frame that are based on inter-worker disagreement. We made a case for this kind of disagreement reflecting genuine cases of ambiguity in FrameNet frames, caused by: child-parent relations between frames, frames with overlapping definitions, or compositions of frames making up the meaning of a word. The evaluation method we proposed uses the scores for multiple frames, and is thus able to differentiate between frames that still apply to the word, but with less clarity, and frames that capture the wrong meaning. Our goal was to build a resource that recognizes different levels of ambiguity in the expression of the frames in the text, and allows a more fair evaluation of performance of frame disambiguation systems.

Acknowledgments

We would like to thank Luigi Asprino, Valentina Presutti and Aldo Gangemi for their assistance with using the Framester corpus, as well as their advice in better understanding the task of frame disambiguation. We would also like to thank the anonymous crowd workers for their contributions to our crowdsourcing tasks.

References

Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation*, 1:31–34.

Collin F Baker. 2012. FrameNet, current collaborations and future goals. *Language Resources and Evaluation*, 46(2):269–286.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Nancy Chang, Praveen Paritosh, David Huynh, and Collin Baker. 2015. Scaling semantic frame annotation. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 1–10.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018a. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018.*, pages 12–20. AAAI Press.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018b. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. *arXiv preprint arXiv:1808.06080*.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Semantic role labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing FrameNet to the crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 742–747.

Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. 2016. Framester: a wide coverage linguistic linked data hub. In *European Knowledge Acquisition Workshop*, pages 239–254. Springer.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1448–1458.

Jisup Hong and Collin F. Baker. 2011. How good is the crowd at "real" WSD? In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 30–37. Association for Computational Linguistics.

Charles J Fillmore. 1982. *Frame Semantics*, volume 34, pages 111–138.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *HLT-NAACL*, pages 556–562.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. FrameNet+: Fast paraphrastic tripling of FrameNet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 408–413.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*.

A Ambiguity Examples in the Corpus

#	SENTENCE	SQS	FRAMES (FSS)
1	These Articles have historically shaped and **continue** to direct the ethos of the Communion.	0.795	*activity ongoing* (0.862) *process continue* (0.86)
2	"A Modest Proposal" is **included** in many literature programs as an example of early modern western satire.	0.771	*inclusion* (0.89) *cause to be included* (0.813)
3	The states often **failed** to meet these requests in full, leaving both Congress and the Continental Army chronically short of money.	0.628	*endeavor failure* (0.826) *success or failure* (0.8)
4	This is a chart of trend of nominal gross domestic product of Angola at market prices **using** International Monetary Fund data.	0.598	*using resource* (0.831) *using* (0.554) *tool purpose* (0.336)
5	The Asian tigers have now all **received** developed country status, having the highest GDP per capita in Asia.	0.504	*receiving* (0.751) *getting* (0.556)
6	MasterCard has released Global Destination Cities Index 2013 with 10 of 20 are **dominated** by Asia and Pacific Region Cities.	0.467	*dominate situation* (0.638) *dominate competitor* (0.579) *being in control* (0.327)

Table 3: Ambiguity because of parent-child relation between frames.

#	SENTENCE	SQS	FRAMES (FSS)
1	Kournikova then **withdrew** from several events due to continuing problems with her left foot and did not return until Leipzig.	0.725	*withdraw from participation* (0.955), *removing* (0.61)
2	Some aikido organizations use belts to **distinguish** practitioners' grades.	0.68	*differentiation* (0.867) *distinctiveness* (0.703)
3	Since then, it has focused on **improving** relationships with Western countries, cultivating links with other Portuguese-speaking countries, and asserting its own national interests in Central Africa.	0.654	*improvement or decline* (0.787) *cause to make progress* (0.732)
4	To **emphasize** the validity of the Levites' claim to the offerings and tithes of the Israelites, Moses collected a rod from the leaders of each tribe in Israel and laid the twelve rods over night in the tent of meeting.	0.65	*emphasizing* (0.764) *convey importance* (0.638)
5	He not only had enough food from his subjects to **maintain** his military, but the taxes collected from traders and merchants added to his coffers sufficiently to fund his continuous wars.	0.453	*cause to continue* (0.7) *activity ongoing* (0.602)
6	He **spent** the later part of his life in the United States, living in Los Angeles from 1937 until his death.	0.29	*taking time* (0.41) *expend resource* (0.365)

Table 4: Ambiguity because of overlapping frame definitions.

#	SENTENCE	SQS	FRAMES (FSS)
1	These writings lack the mystical, philosophical elements of alchemy, but do contain the works of Bolus of Mendes (or Pseudo-Democritus), which **aligned** these recipes with theoretical knowledge of astrology and the classical elements.	0.284	*arranging* (0.474) *adjusting* (0.4) *assessing* (0.298) *compatibility* (0.254) *undergo change* (0.169)
2	However, commercial application of this fact has challenges in **circumventing** the passivating oxide layer, which inhibits the reaction, and in storing the energy required to regenerate the aluminium metal.	0.239	*dodging* (0.477) *compliance* (0.248) *surpassing* (0.204) no frame (0.148)
3	This had the effect of **inculcating** the principle of "Lex orandi, lex credendi" (Latin loosely translated as 'the law of praying [is] the law of believing') as the foundation of Anglican identity and confession.	0.201	*education teaching* (0.384) *communication* (0.35) no frame (0.153)
4	Legal segregation ended in the states in 1964, but Jim Crow customs often continued until specifically **challenged** in court.	0.172	*difficulty* (0.372) *competition* (0.283) *taking sides* (0.257) *communication* (0.154)
5	When Washington's army arrived outside Yorktown, Cornwallis prematurely abandoned his outer position, **hastening** his subsequent defeat.	0.134	*speed description* (0.39) *assistance* (0.209) *self motion* (0.165) *travel* (0.16) *causation* (0.124)

Table 5: Ambiguity because the meaning of the word is expressed by a composition of frames.