Uncertainty Detection for Natural Language Watermarking

György Szarvas[1] **Iryna Gurevych**[1,2]
(1) Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
(2) Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de

Abstract

In this paper we investigate the application of uncertainty detection to text watermarking, a problem where the aim is to produce individually identifiable copies of a source text via small manipulations to the text (e.g. synonym substitutions). As previous attempts showed, accurate paraphrasing is challenging in an open vocabulary setting, so we propose the use of the closed word class of uncertainty cues. We demonstrate that these words are promising for text watermarking as they can be accurately disambiguated (from the non-cue uses of the same words) and their substitution with other cues has marginal impact to the meaning of the text.

1 Introduction

The goal of digital watermarking is to hide digital information (a secret marker) in an audio stream, image or text file. These markers are by design not perceivable while listening, watching or reading the data, but can be read with a tailor-made algorithm and can be used to authenticate the data that carries it, or to identify its owner. We discuss the concept of a watermark and the process of embedding it in a media in more detail in Section 2. Text watermarking is a digital watermarking problem where the aim is to embed a secret message (a sequence of bits) in a *text* in order to make the actual text copy individually identifiable (Bergmair, 2007). That is, given a natural language source text (e.g. an ebook), the aim is to produce individual copies of the text by means of surface, syntactic or semantic manipulations. The individualized copies should preserve the readability and the meaning of the original text, i.e. the modifications should be undetectable to the readers.

The manipulation of the text can be performed on the surface or the content level. Surface manipulation affects the visual appearance of the text, as in white space modulation. In contrast, syntactic reordering looks e.g. for conjunctions and reorders the connected words or phrases. Finally, semantic manipulation, such as *synonym substitution*, takes a target word and replaces it with a contextually plausible synonym. Consider the following examples:

He was *bright*␣and independent and proud.
He was *bright*␣␣and independent and proud. (surface)
He was independent and *bright* and proud. (syntactic)
He was *clever* and independent and proud. (semantic)

In steganography, there is a natural tradeoff between capacity (the length of the secret message that can be embedded in a cover medium) and precision (how well the manipulations preserve readability and meaning). Text watermarking is a precision- oriented application as it requires very high transformation quality, and relatively low capacity (recall) is acceptable as the goal is to make a relatively small number of changes in a text. Surface manipulations are typically very easy to spot, so those are not realistic alternatives, while syntactic and semantic methods are equally viable. From a practical perspective, the most reliable method should be used to embed the secret information. E.g. enumeration reordering is a relatively robust method, provided we have a good parser at hand to detect the conjoined units. On the other hand, if there is a reference in the near context, it can be more reliable to employ synonym substitution:

He made offers to John and Mary. The latter *accepted*.
He made offers to John and Mary. The latter *agreed*.

Here we investigate paraphrasing as a way to embed secret information in texts. An open vocabulary approach to synonym substitution could ensure rather high capacity. For example, Chang and Clark (2010b) suggest to identify one paraphrase position per sentence, thereby enabling the system to encode one bit per sentence, but they achieve

International Joint Conference on Natural Language Processing, pages 1188–1194,
Nagoya, Japan, 14-18 October 2013.

original bitmap	#	after embedding	#
1 1 1 0 0 0	1	1 1 1 0 0 0	1
1 1 1 1 0 1	1	1 1 1 0 0 1	0
1 0 1 0 0 1	1	1 0 1 1 0 1	0
1 1 0 0 0 1	1	1 1 0 0 0 1	1
1 0 0 0 0 1	0	1 0 0 0 0 1	0
0 0 0 0 1 1	0	0 0 0 0 1 1	0

Table 1: Example watermark.

this at the cost of relatively low precision (around 70% even when using the information about the correct word sense). In contrast, we propose to use a closed word class – i.e. uncertainty cues – for paraphrasing. Semantically uncertain propositions are common in natural language: they either directly express something as uncertain (epistemic modality); assert the speaker's hypotheses, beliefs; express events that are unconfirmed or under investigation or conditional to other events, etc. These linguistic devices are lexical in nature, i.e. they are triggered by the use of specific keywords in the sentence, which we refer to as *uncertainty cues*. These words are good targets for paraphrasing since – when replaced by another uncertainty cue with similar properties (part of speech and meaning) – their substitution does not change the meaning of the sentence: the main proposition ("who does what, when and where") remains the same and the proposition stays uncertain:

The legislative body *may* change.
The legislative body *might* change.

That is, in our approach a classifier first detects uncertainty cues in a text. Then, those disambiguated cues that are both detected with high precision and have such paraphrases that are valid in all "uncertain" contexts can be replaced with their paraphrase. This manipulation affects the watermark bit contained in the passage and thus allows information encoding. The actual process is described in more detail in Section 2.1.

With this work, we show that a closed-class approach to generating paraphrases has desirable properties for watermarking: almost perfect substitution quality and acceptable embedding capacity. At the same time, we propose a novel application of uncertainty cue detection: paraphrasing of uncertainty cues for linguistic steganography.

2 Digital Watermarking

In this section we elaborate on what a watermark message is. In the simplest setup, we can define a watermark message as a sequence of k bits (0 or 1 values). Then the message then can take 2^k different values, and can be used to identify the owner of the medium, if each owner gets a copy of the data with a different digital message embedded in it. Take, for example, the problem of embedding a 6-bit message in a black-and-white bitmap image. In this case, we divide the bitmap to six equal-size parts and consider the parity of the sum of bits in a given part as the message bit. Table 1 shows a 6×6 bitmap and the embedding of a 6-bit message (100100, one bit in each line). For a comprehensive overview of the different watermarking techniques, we refer the reader to Cox et al. (2008).

2.1 Natural Language Watermarking with Paraphrases

In order to encode a single bit in a larger block of text – like a paragraph or section – based on bit parity, we first assign a bit (0 or 1) to every word in the vocabulary, i.e. not just those that we plan to manipulate. Then, in each block of the source text, we sum the bits encoded by the original text and take its parity ($even = 0$, $odd = 1$) as the secret bit. In case our message should contain the other bit, we have to make exactly 1 synonym substitution, replacing a 0-word with a 1-word or vice versa to reverse the parity of the block, in order to embed the desired bit in the text. Reading the message only requires the same methodology to detect the blocks (sections) in the text and the initial word-to-bit mapping to calculate the parity of the blocks.

If we consider only a small set of words as candidates for substitution, this places a constraint on the system: it has to be ensured that the receiver can identify which blocks are used to actually encode information. This can be done, for example, if we use such blocks that contain at least one paraphrasable word after the potential manipulation: this way the reader can be sure the actual block encodes a bit (as there is capacity left in the block).

This simple parity-based message encoding approach offers straightforward ways to combine different embedding techniques, which is desirable: in order to combine conjunction reordering with the proposed paraphrasing method, one could use the (parity of the) number of changes in the conjoined list of items from their lexicographic order as the bit encoded in a conjunction, and can use that bit in the message encoding process. In our application scenario, we plan to implement several different text manipulation methods (and combine

them based on their confidence). Therefore the goal of this study is to propose an approach with high precision and coverage will be ensured by a range of syntactic methods and paraphrasing together.

3 Related Work

In this section we briefly introduce the previous work in *natural language watermarking* and in *uncertainty detection*.

The most prominent previous work in natural language watermarking (Atallah et al., 2003) focus on manipulating the sentence structure in order to realize meaning-preserving transformations of the text, which in turn can be used to embed information in the text. This approach either requires the presence of a robust syntactic (and semantic) parser to construct the representation which supports complex sentence-level transformations, or it has to be simplified to local transformations (e.g. conjunction modulation, as in the examples above) in order to ensure high precision without the requirement of deep linguistic analysis. Unfortunately, robust syntactic and semantic parsing of arbitrary texts is still challenging for natural language processing. This fact justifies the importance of more shallow models, such as synonym substitution, provided these approaches can ensure high precision and robust performance across domains. For a general and detailed overview of linguistic steganography, including methods other than paraphrasing, see for example Bennett (2004) and Topkara et al. (2005).

3.1 Paraphrasing for Linguistic Steganography

As regards synonym substitution, the first studies made no use (Topkara et al., 2006) or just limited use (Bolshakov, 2005) of the context through collocation lists. While this approach offers a relatively high capacity, the transformations result in frequent semantic, or even grammatical, errors in the text, which is undesirable (Bennett, 2004).

Recently Chang and Clark (2010a,b) proposed the use of contextual paraphrases for linguistic steganography. This offers a higher perceived quality and is therefore more suited to text watermarking where quality is a crucial aspect. Chang and Clark (2010b) used the English Lexical Substitution data (McCarthy and Navigli, 2007) from SemEval 2007 for evaluation, WordNet as the source of potential synonyms and n-gram frequencies for candidate ranking to experiment with paraphrasing for linguistic steganography. They introduced a graph coloring approach to embed information in a text through the substitution of words with their WordNet synonyms. They report an accuracy slightly above 70% for their paraphrasing technique and a potential capacity of around one bit per sentence.

Chang and Clark (2010a) used a paraphrase dictionary mined from parallel corpora using statistical machine translation (SMT) methods. The ranking of candidate paraphrases was also based on n-gram frequencies and for a set of 500 paraphrase examples they reported 100% precision (with very low, 4% recall) for substitution grammaticality. The possible change in meaning for otherwise grammatical replacements was not evaluated.

3.2 Uncertainty Cue Detection

Another major field of work related to this study is the detection of uncertainty cues, which we propose to use for paraphrasing in Section 4. The first approaches to uncertainty detection were based on hand-crafted lexicons (Light et al., 2004; Saurí, 2008). In particular, ConText (Chapman et al., 2007) used lexicons and regular expressions not only to detect cues, but also to recognize contexts where a cue word does not imply uncertainty.

Supervised uncertainty cue detectors have also been developed using either token classification (Morante and Daelemans, 2009) or sequence labeling approaches (Tang et al., 2010). A good overview and comparison of different statistical approaches is given in Farkas et al. (2010). Szarvas et al. (2012) addressed uncertainty cue detection in a multi-domain setting, using surface-level, part-of-speech and chunk-level features and sequence labeling (CRF) models. They found that cue words can be accurately detected in texts with various topics and stylistic properties. We make use of the multidomain corpora presented in their study and evaluate a cross-domain cue detection model for text watermarking.

4 Uncertainty Cue Detection for Text Watermarking

In this section we experiment with uncertainty cue detection and paraphrasing, and study the potential of this approach for text watermarking.

4.1 Dataset, Experimental Setup and Evaluation Measures

We used here the dataset introduced by Szarvas et al. (2012). It consists of texts from three different domains (articles from Wikipedia, newspaper articles and biological scientific texts) and is annotated for uncertainty cues. The uncertainty cues in the corpora are marked only in contexts where they are used in an uncertain meaning, i.e. these texts can be used to train a shallow (`cue` vs. `non-cue` meaning) disambiguation model for these phrases. Here we aim to paraphrase the `cue` uses of the words to encode information via cue-to-cue paraphrasing. We train and test our models on separate parts of the corpora: for example, to assess the accuracy of cue detection in Wikipedia texts, we train the model only on newswire and scientific texts, and so on. This is a cross-domain evaluation setting and is therefore a good estimate of how the system would perform on further, yet different text types from our training datasets.

For evaluation, we use the overall precision of the recognized uncertainty cues, and we also measure the capacity that can be achieved by paraphrasing these words, i.e. how frequently one of these words that we use to encode information actually occurs in a text (the number of detected objects divided by the number of sentences processed). These two criteria – precision and capacity – measure how well the uncertainty detector would perform as a stego system. In addition, we also perform an error analysis of the top-ranked instances that received the highest posterior probability scores by the classifier. The highest-ranked instances are especially important as the underlying application would chose the highest-ranked instance in a larger block of text to actually implement a change to the text.

4.2 Uncertainty Detection Model

We implemented a cue recognition model similar to that described in Szarvas et al. (2012), using simple features that are robust across various text types. This is important, as we plan to use the model for text types that can be different from those in the training corpus, and for which NLP modules such as parsers might have questionable performance.

Conditional Random Fields were found to provide the best overall results in cue detection. However, the relative advantage of sequence taggers over simple token-based classifiers is more prominent for less frequent, long cue phrases[1]. Since in this study we concentrate on the more simple and frequent unigram cues (or fixed constructions, such as *not clear*), we use a Maximum Entropy (Maxent) classifier model (McCallum, 2002) in our experiments for cue detection and disambiguation.

In our classification setup, each token is a separate instance, described by a set of features collected from a window of size 2 around the token. That is, each feature can be present under 5 different names, according to their relative position, except for sentence beginning and ending tokens (out-of-sentence positions are discarded). We used the following features to describe each token: i) lexical features (word base forms and word bigrams); ii) 3 to 5 characters long word prefixes and suffixes; iii) word-surface-level features denoting whether the word is written in all uppercase, uppercase initial, or lowercase form, it is a punctuation mark or number; and iv) part of speech tags.

corpus	# sent.	# cues	precision	capacity	F(cue)
Wiki	20756	3438	69.69%	16.56%	71.28%
news	3123	522	84.48%	16.71%	70.33%
sci	19473	3515	91.58%	18.05%	70.92%

Table 2: Summary of cue recognition results.

4.3 Results

The results of the cross-domain cue recognition experiments are summarized in Table 2. The columns indicate the total number number of sentences and recognized cues in the corpora, their precision and the capacity that can be achieved via cue paraphrasing. For comparison to previous works, we also provide the overall phrase-level F score for uncertainty cues. These numbers are slightly better than those reported by Szarvas et al. (2012) for cross-training with CRFs, probably due to the different settings (we used two domains for training, not just one).

As can be seen, the uncertainty cue recognizer is accurate even in a challenging cross-training setting: the precision is well above 80% for two out of three domains, and is around 70% for the most difficult Wikipedia dataset. This precision could realize a capacity of one bit per every six sentences, on average (capacity at or above 16%). In

[1] We performed an initial experiment using token-based and sequential models on our corpora and found no statistically significant difference in performance on unigram cues.

corpus	# sent.	# cues	precision	capacity
Wiki	20756	1869	89.46%	9.00%
news	3123	223	93.72%	7.14%
sci	19473	2688	98.95%	13.80%

Table 4: Summary of results with the 29 selected keywords.

order to use this cue recognizer as a watermarking method with the above-mentioned precision and capacity, we should provide a valid paraphrase for all of the 300 uncertainty cues found in the corpora. Doing that, a precision of 70% or more is promising in the light of the precision values below 50% for the first answers at SemEval 2007 for an all-words substitution task (Yuret, 2007), and of the fact that this precision stays around 70% even if the correct sense is picked in advance based on the human answers (Chang and Clark, 2010b).

On the other hand, as we argued in Section 2.1, it is desirable to improve precision at the cost of capacity. Thus, we filtered the uncertain vocabulary for such cues that are both frequent and accurate: we kept the cues that had a frequency of at least 10, with a precision above 80%. This left us 37 cues in total and this list was given to two annotators to provide paraphrases. The annotators were told to perform a web search for various contexts of the words and suggest paraphrases that are acceptable in all the words' uncertainty-cue-uses and contexts. The two annotators agreed on a unique paraphrase which fits in all uncertain contexts of the target words for only 29 cues. These words with the proposed paraphrases and examples from the Wikipedia corpus are listed in Table 3. The columns indicate the selected cue words with their part of speech and the proposed paraphrase cue (or *XX* where we could not provide a suitable paraphrase). As can be seen, each cue is paraphrased with another cue with the same part of speech. Thus, their inflected forms can be generated based on the original words' POS tags. For the remaining eight cues the annotators either did not find a proper substitute (e.g. *belief*) or found the word to have several uncertain readings which would require different paraphrases in different contexts (e.g. *expect* which can be rephrased as *wait, hope, count on, ...*).

Table 4 provides aggregate numbers for the selected 29 cue words. The columns indicate the total number number of sentences in the corpora, the number of recognized instances of the 29 selected cues and their precision and capacity. As can be seen, for these words the classifier yielded excellent precision scores, even with cross-domain training. In scientific and newswire texts, the model performs well above 90% precision, while in Wikipedia texts the precision is slightly lower.

As regards the capacity of the selected cues in the texts, on average we can find one instance of the selected cue words in every 7–14 sentences. The above precision and capacity scores can be realized in an actual watermarking method with the use of the paraphrases in Table 3. While this coverage seems lower than some other approaches (e.g. Chang and Clark (2010b) can achieve approximately one bit per sentence capacity), it is still acceptable for text watermarking of lengthy documents (such as ebooks), and as mentioned earlier, different methods can be combined to increase capacity. In the light of this, we consider our results especially promising, due to the remarkable precision scores, and the positive characteristic that these changes do not affect the main propositions in the sentences, i.e. meaning is well preserved.

Although direct comparisons are difficult to make, Chang and Clark (2010a) evaluated how accurately their model predicted a paraphrase to be grammatical, which is similar to our goal here. Their model achieved similar precision levels with similar or slightly lower potential capacity scores. Other previous approaches reported substantially lower precision (typically aiming to achieve high recall). These results suggest that our methodology, making a change in 7–14% of the sentences with a precision of 90–98% is a very competitive alternative for precision-oriented linguistic steganography.

4.4 Error Analysis

The proposed method can embed information in a cover text with remarkably high precision, as the applied changes to the text are perfectly grammatical and do not affect the main aspects of the meaning of the text. Our error analysis also confirms this (see Appendix). We checked the 250 top-ranked classifications in the Wikipedia and scientific text corpora, and 223 classifications in the newswire texts (the total number of detected instances of the 29 selected cues). We checked the errors in the instances that obtained the highest posterior scores because in a larger block the sys-

WORD	POS	SUBST.	Example
accuse (of)	V	blame (for)	Certain corporations have been **accused** of paying news channels.
allege	V	claim	A friend of his **alleges** it detects ghosts.
allegedly	RB	reportedly	Britain was **allegedly** fighting for the freedom of Europe.
assume	V	hypothesize	It is **assumed** that women are not capable of inflicting such harm.
assumption	N	hypothesis	They respond that the **assumption** has long been that he worked from a sketch.
belief	N	XX	It was common **belief** that all species came to existence by divine creation.
believe	V	think	These were **believed** to be in the CA $150.000 range.
determine	V	XX	... but ongoing studies have **yet to determine** to what degree.
expect	V	XX	He **expects** to be promoted to a grade 35 bureaucrat.
hypothesize	V	assume	It is **hypothesized** that most of these chemicals help.
hypothesis	N	assumption	There is some evidence to support the **hypothesis** that they udergo fission.
idea	N	XX	The **idea** that it constitutes an edifice was publicized by Osmanagic.
imply	V	denote	It is **implied** to be the center of the Dust Factory.
indicate	V	suggest	It **indicates** that there are good opportunities for skilled people.
likely	JJ	probable	It is **likely** that they were instructed by their grandmother M. V. van Aelst.
likely	RB	probably	The camps will **likely** never reopen as their locations posed lightning risks.
may	MD	might	The legislative body **may** change or repeal any prior legislative acts.
might	MD	may	Other instruments that **might** be connected are air data computers.
not clear	JJ	unclear	How the plant arrived on the island is **not clear**.
perhaps	RB	maybe	His work was **perhaps** known to Islamic mathematicians.
possibility	NN	potential	However the **possibility** of merging University Park with Downtown LA remains years away.
possibly	RB	potentially	It is **possibly** a close relative to the dwarf flannelbush species.
presumably	RB	supposedly	Wellstone was **presumably** worried about money from rich individuals.
probably	RB	likely	He was **probably** better known for his antics than his pitching talent.
regard	V	XX	Shea Fahy is **regarded** as one of the heroes of the team.
seem	V	appear	It **seems** that Kev takes the opportunity to ...
seemingly	RB	apparently	Pelham was **seemingly** intimate with John Smibert.
speculate	V	assume	Some people **speculate** that these compounds are linked to health concerns.
suggest	V	indicate	It **suggests** that those few cases have their needs already met.
suppose	V	assume	The "arms" of the bow are **supposed** to cross each other.
suspect	V	XX	The diagnosis is often **suspected** on the basis of tests.
think	V	XX	Most people did not think that the Rams belonged on the same field with the Steelers.
thought	V	XX	Sleep is **thought** to improve the consolidation of information.
unclear	JJ	not clear	It is **unclear** whether it was House or Wilson.
unlikely	JJ	not likely	Historians are **unlikely** to fully understand which species were used in medicine.
view (that)	N	opinion (that)	... that undermined their capacity to accept the **view** that socialist incentives would not work
whether	IN	if	... or **whether** his paintings were purchased by Italians.

Table 3: Uncertain paraphrase dictionary with examples from the Wikipedia corpus.

tem would select the most confident position to perform a substitution, so precision at top ranks is the most important. We found 20 false positive classifications in these 723 sentences, attributed to eight different keywords. This is above 97% precision at top ranks, the errors are detailed in the appendix, together with example sentences. As can be seen, many of these misclassifications actually do not do any major harm to the meaning of the text: some of these high-ranked examples are actually replaceable with the proposed cover words even in a non-cue usage (this is the reason for their high posterior score).

Conclusion and Future Work

In this paper we proposed uncertainty cue detection and the paraphrasing of uncertainty cues as a new approach to linguistic steganography. We experimented with texts from three different domains using cue detection models trained on out-of-domain texts in order to simulate a realistic application scenario. We found that uncertainty cues are capable of embedding a 1-bit message in a block of text of around 10–13 sentences on average. Although this capacity is limited, in turn the use of uncertainty cues offers nearly perfect precision, i.e. the manipulated texts are grammatical and preserve the original text's meaning. As in text watermarking the goal is to embed a relatively short message in potentially large texts, but with high precision (quality), the paraphrasing of uncertainty cues is a promising alternative.

Arguably, the ideal setting from an application perspective would be an open-vocabulary substitution system, but such an approach suffers from significant limitations stemming from the difficulty of paraphrasing in a general setting. An alternative could be to use a larger set of frequent words in a lexicalized approach for lexical substitution which might offer higher capacity with comparable precision Biemann and Nygaard (2010).

On the other hand, we think it is a viable approach to target the extra-propositional aspects of meaning (such as uncertainty proposed here) for lexical substitution in linguistic steganography. This – by definition – leaves the main proposition of the sentence (*who does what, when and where*) untouched, ensuring high transparency. To this end, we also plan to extend the capacity of this approach via paraphrasing other word classes such as opinion expressions (e.g. *great X* for *excellent X*, *awful X* for *terrible X*).

Appendix

WORD	#	SUBST.
suggest	2	indicate
may	6	might
might	2	may
likely	1	probable
believe	2	think
assume	3	hypothesize
indicate	3	suggest
possibility	1	potential

Table 5: Examples for the 21 errors in the top 250 (wiki, sci) and 223 (news) examples.

Example Errors

– Churchill wrote to him **suggesting** that he would sign his own works "Winston S. Churchill".
– He **may** be an idiot savant, but he's got great hair.
– It's fairly intense as you **might** well imagine.
– One big question now is the **likely** role of Mr. Fournier's allies.
– Nobody **believe** this any more.
– Cilcior will also **assume** 22 million of Hunter's existing debt.
– Kellogg **indicated** that it has room to grow without adding facilities.
– The second **possibility** would be to start a fight with Israel.

Acknowledgment

This work was supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program (I/82806), and by the German Ministry of Education and Research under grant *SiDiM* (grant No. 01IS10054G).

References

M. J. Atallah, V. Raskin, C. Hempelmann, M. Karahan, R. Sion, U. Topkara, and K. E. Triezenberg. 2003. Natural language watermarking and tamperproofing. In *Proc. of 5th Workshop on Information Hiding*, pages 196–212.

K. Bennett. 2004. Linguistic steganography: Survey, analysis, and robustness concerns for hiding information in text. Technical report, Purdue University.

R. Bergmair. 2007. A comprehensive bibliography of linguistic steganography. In *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents*.

C. Biemann and V. Nygaard. 2010. Crowdsourcing wordnet. In *Proceedings of the 5th Global WordNet conference*.

I. Bolshakov. 2005. A method of linguistic steganography based on collocationally-verified synonymy. In *Information Hiding*, volume 3200 of *LNCS*, pages 607–614.

C-Y. Chang and S. Clark. 2010a. Linguistic steganography using automatically generated paraphrases. In *Proceedings of NAACL 2010*, pages 591–599.

C-Y. Chang and S. Clark. 2010b. Practical linguistic steganography using contextual synonym substitution and vertex colour coding. In *Proceedings of the EMNLP 2010*, pages 1194–1203.

W. W. Chapman, D. Chu, and J. N. Dowling. 2007. ConText: An algorithm for identifying contextual features from clinical text. In *Proceedings of BioNLP 2007*, pages 81–88.

I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker. 2008. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., 2 edition.

R. Farkas, V. Vincze, Gy. Móra, J. Csirik, and Gy. Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of CoNLL: Shared Task*, pages 1–12.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

H. Kilicoglu and S. Bergler. 2008. Recognizing speculative language in biomedical research articles: A linguistically motivated perspective. In *Proceedings of the BioNLP Workshop*, pages 46–53.

M. Light, X. Y. Qiu, and P. Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of Biolink 2004 Ws.*, pages 17–24.

A. K. McCallum. 2002. *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

D. McCarthy and R. Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of SemEval-2007*, pages 48–53.

B. Medlock and T. Briscoe. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of ACL*, pages 992–999.

R. Morante and W. Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP 2009 Workshop*, pages 28–36.

R. Saurí. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University, Waltham, MA.

Gy. Szarvas, V. Vincze, R. Farkas, Gy. Móra, and I. Gurevych. 2012. Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. *Computational Linguistics*, 38(2):335–367.

B. Tang, X. Wang, X. Wang, B. Yuan, and S. Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of CoNLL-2010: Shared Task*, pages 13–17.

M. Topkara, C. M. Taskiran, and E. J. Delp. 2005. Natural language watermarking. In *Security, Steganography, and Watermarking of Multimedia Contents*, pages 441–452.

U. Topkara, M. Topkara, and M. J. Atallah. 2006. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th workshop on Multimedia and security*, pages 164–174.

D. Yuret. 2007. Ku: Word sense disambiguation by substitution. In *Proceedings of SemEval-2007*, pages 207–214.