Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features

Viola Ganter and **Michael Strube**
EML Research gGmbH
Heidelberg, Germany
http://www.eml-research.de/nlp

Abstract

We investigate the automatic detection of sentences containing linguistic hedges using corpus statistics and syntactic patterns. We take Wikipedia as an already annotated corpus using its tagged weasel words which mark sentences and phrases as non-factual. We evaluate the quality of Wikipedia as training data for hedge detection, as well as shallow linguistic features.

1 Introduction

While most research in natural language processing is dealing with identifying, extracting and classifying facts, recent years have seen a surge in research on sentiment and subjectivity (see Pang & Lee (2008) for an overview). However, even opinions have to be backed up by facts to be effective as arguments. Distinguishing facts from fiction requires to detect subtle variations in the use of linguistic devices such as linguistic hedges which indicate that speakers do not back up their opinions with facts (Lakoff, 1973; Hyland, 1998).

Many NLP applications could benefit from identifying linguistic hedges, e.g. question answering systems (Riloff et al., 2003), information extraction from biomedical documents (Medlock & Briscoe, 2007; Szarvas, 2008), and deception detection (Bachenko et al., 2008).

While NLP research on classifying linguistic hedges has been restricted to analysing biomedical documents, the above (incomplete) list of applications suggests that domain- and language-independent approaches for hedge detection need to be developed. We investigate Wikipedia as a source of training data for hedge classification. We adopt Wikipedia's notion of *weasel words* which we argue to be closely related to hedges and private states. Many Wikipedia articles contain a specific *weasel tag*, so that Wikipedia can be viewed as a readily annotated corpus. Based on this data, we have built a system to detect sentences that contain linguistic hedges. We compare a baseline relying on word frequency measures with one combining word frequency with shallow linguistic features.

2 Related Work

Research on hedge detection in NLP has been focused almost exclusively on the biomedical domain. Light et al. (2004) present a study on annotating hedges in biomedical documents. They show that the phenomenon can be annotated tentatively reliably by non-domain experts when using a two-way distinction. They also perform first experiments on automatic classification.

Medlock & Briscoe (2007) develop a weakly supervised system for hedge classification in a very narrow subdomain in the life sciences. They start with a small set of seed examples known to indicate hedging. Then they iterate and acquire more training seeds without much manual intervention (step 2 in their seed generation procedure indicates that there is some manual intervention). Their best system results in a 0.76 precision/recall break-even-point (BEP). While Medlock & Briscoe use words as features, Szarvas (2008) extends their work to n-grams. He also applies his method to (slightly) out of domain data and observes a considerable drop in performance.

3 Weasel Words

Wikipedia editors are advised to avoid *weasel words*, because they "offer an opinion without really backing it up, and ... are really used to express a non-neutral point of view."[1] Examples for weasel words as given by the style guide-

[1] http://en.wikipedia.org/wiki/Wikipedia:Guide_to_writing_better_articles

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 173–176,
Suntec, Singapore, 4 August 2009.

lines[2] are: *"Some people say ... ", "I think ... ", "Clearly ... ", "...is widely regarded as ... ", "It has been said/suggested/noticed ... ", "It may be that ... "* We argue that this notion is similar to linguistic hedging, which is defined by Hyland (1998) as "...any linguistic means used to indicate either a) a lack of complete commitment to the truth value of an accompanying proposition, or b) a desire not to express that commitment categorically." The Wikipedia style guidelines instruct editors to, if they notice weasel words, insert a {{`weasel-inline`}} or a {{`weasel-word`}} tag (both of which we will hereafter refer to as weasel tag) to mark sentences or phrases for improvement, e.g.

(1) `Others argue {{weasel-inline}} that the news media are simply catering to public demand.`

(2) `...therefore America is viewed by some {{weasel-inline}} technology planners as falling further behind Europe ...`

4 Data and Annotation

Weasel tags indicate that an article needs to be improved, i.e., they are intended to be removed after the objectionable sentence has been edited. This implies that weasel tags are short lived, very sparse and that – because weasels may not have been discovered yet – not all occurrences of linguistic hedges are tagged. Therefore we collected not one but several Wikipedia dumps[3] from the years 2006 to 2008. We extracted only those articles that contained the string {{`weasel`. Out of these articles, we extracted 168,923 unique sentences containing 437 weasel tags.

We use the dump completed on July 14, 2008 as development test data. Since weasel tags are very sparse, any measure of precision would have been overwhelmed by false positives. Thus we created a balanced test set. We chose one random, non-tagged sentence per tagged sentence, resulting (after removing corrupt data) in a set of 500 sentences. We removed formatting, comments and links to references from all dumps. As testing data we use the dump completed on March 6, 2009. It comprises 70,437 sentences taken from articles containing the string {{`weasel` with 328 weasel tags. Again, we created a balanced set of 500 sentences.

	S	M	C
K	0.45	0.71	0.6
S		0.78	0.6
M			0.8

Table 1: Pairwise inter-annotator agreement

As the number of weasel tags is very low considering the number of sentences in the Wikipedia dumps, we still expected there to be a much higher number of potential weasel words which had not yet been tagged leading to false positives. Therefore, we also annotated a small sample manually. One of the authors, two linguists and one computer scientist annotated 100 sentences each, 50 of which were the same for all annotators to enable measuring agreement. The annotators labeled the data independently and following annotation guidelines which were mainly adopted from the Wikipedia style guide with only small adjustments to match our pre-processed data. We then used *Cohen's Kappa* (κ) to determine the level of agreement (Carletta, 1996). Table 4 shows the agreement between each possible pair of annotators. The overall inter-annotator agreement was $\kappa = 0.65$, which is similar to what Light et al. (2004) report but worse than Medlock & Briscoe's (2007) results. As Gold standard we merged all four annotations sets. From the 50 overlapping instances, we removed those where less than three annotators had agreed on one category, resulting in a set of 246 sentences for evaluation.

5 Method

5.1 Words Preceding Weasel Tags

We investigate the five words occurring right before each weasel tag in the corpus (but within the same sentence), assuming that weasel phrases contain at most five words and weasel tags are mostly inserted *behind* weasel words or phrases.

Each word within these 5-grams receives an individual score, based a) on the relative frequency of this word in weasel contexts and the corpus in general and b) on the average distance the word has to a weasel tag, if found in a weasel context. We assume that a word is an indicator for a weasel if it occurs close before a weasel tag. The final scoring function for each word in the training set

[2] `http://en.wikipedia.org/wiki/Wikipedia:Avoid_weasel_words`

[3] `http://download.wikipedia.org/`

is thus:

$$Score(w) = RelF(w) + AvgDist(w) \tag{1}$$

with

$$RelF(w) = \frac{W(w)}{log_2(C(w))} \tag{2}$$

and

$$AvgDist(w) = \frac{W(w)}{\sum_{j=0}^{W(w)} dist(w, weaseltag_j)} \tag{3}$$

$W(w)$ denotes the number of times word w occurred in the context of a weasel tag, whereas $C(w)$ denotes the total number of times w occurred in the corpus. The basic idea of the $RelF$ score is to give those words a high score, which occur frequently in the context of a weasel tag. However, due to the sparseness of tagged instances, words that occur with a very high frequency in the corpus automatically receive a lower score than low-frequent words. We use the logarithmic function to diminish this effect.

In equation 3, for each weasel context j, $dist(w, weaseltag_j)$ denotes the distance of word w to the weasel tag in j. A word that always appears directly before the weasel tag will receive an $AvgDist$ value of 1, a word that always appears five words before the weasel tag will receive an $AvgDist$ value of $\frac{1}{5}$. The score for each word is stored in a list, based on which we derive the classifier (*words preceding weasel (wpw)*): Each sentence S is classified by

$$S \rightarrow weasel\ \ if\ \ wpw(S) > \sigma \tag{4}$$

where σ is an arbitrary threshold used to control the precision/recall balance and $wpw(S)$ is the sum of scores over all words in S, normalized by the hyperbolic tangent:

$$wpw(S) = \tanh \sum_{i=0}^{|S|} Score(w_i) \tag{5}$$

with $|S|$ = the number of words in the sentence.

5.2 Adding shallow linguistic features

A great number of the weasel words in Wikipedia can be divided into three categories:

1. Numerically underspecified subjects (*"Some people", "Experts", "Many"*)
2. Passive constructions (*"It is believed", "It is considered"*)
3. Adverbs (*"Often", "Probably"*)

We POS-tagged the test data with the TnT tagger (Brants, 2000) and developed finite state automata to detect such constellations. We combine these syntactic patterns with the word-scoring function from above. If a pattern is found, only the head of the pattern (i.e., adverbs, main verbs for passive patterns, nouns and quantifiers for numerically underspecified subjects) is assigned a score. The scoring function *adding syntactic patterns (asp)* for each sentence is:

$$asp(S) = \tanh \sum_{i=0}^{heads_S} Score(w_i) \tag{6}$$

where $heads_S$ = the number of pattern heads found in sentence S.

6 Results and Discussion

Both, the classifier based on *words preceding weasel (wpw)* and the one based on *added syntactic patterns (asp)* perform comparably well on the development test data. *wpw* reaches a 0.69 precision/recall break-even-point (BEP) with a threshold of σ = 0.99, while *asp* reaches a 0.70 BEP with a threshold of σ = 0.76.

Applied to the test data these thresholds yield an F-Score of 0.70 for *wpw* (prec. = 0.55/rec. = 0.98) and an F-score of 0.68 (prec. = 0.69/rec. = 0.68) for *asp* (Table 2 shows results at a few fixed thresholds allowing for a better comparison). This indicates that the syntactic patterns do not contribute to the regeneration of weasel tags. Word frequency and distance to the weasel tag are sufficient.

The decreasing precision of both approaches when trained on more tagged sentences (i.e., computed with a higher threshold) might be caused by the great number of unannotated weasel words. Indeed, an investigation of the sentences scored with the added syntactic patterns showed that many high-ranked sentences were weasels which had not been tagged. A disadvantage of the weasel tag is its short life span. The weasel tag marks a phrase that needs to be edited, thus, once a weasel word has been detected and tagged, it is likely to get removed soon. The number of tagged sentences is much smaller than the actual number of weasel words. This leads to a great number of false positives.

σ	.60	.70	**.76**	.80	.90	**.98**
balanced set						
wpw	.68	.68	.68	.69	.69	**.70**
asp	.67	.68	**.68**	.68	.61	.59
manual annot.						
wpw	-	.59	-	-	-	**.59**
asp	.68	.69	**.69**	.69	.70	.65

Table 2: F-scores at different thresholds (bold at the precision/recall break-even-points determined on the development data)

The difference between *wpw* and *asp* becomes more distinct when the manually annotated data form the test set. Here *asp* outperforms *wpw* by a large margin, though this is also due to the fact that *wpw* performs rather poorly. *asp* reaches an F-score of 0.69 (prec. = 0.61/rec. = 0.78), while *wpw* reaches only an F-Score of 0.59 (prec. = 0.42/ rec. = 1). This suggests that the added syntactic patterns indeed manage to detect weasels that have not yet been tagged.

When humans annotate the data they not only take specific words into account but the whole sentence, and this is why the syntactic patterns achieve better results when tested on those data. The word frequency measure derived from the weasel tags is not sufficient to cover this more intelligible notion of hedging. If one is to be restricted to words, it would be better to fall back to the weakly supervised approaches by Medlock & Briscoe (2007) and Szarvas (2008). These approaches could go beyond the original annotation and learn further hedging indicators. However, these approaches are, as argued by Szarvas (2008) quite domain-dependent, while our approach covers the entire Wikipedia and thus as many domains as are in Wikipedia.

7 Conclusions

We have described a hedge detection system based on word frequency measures and syntactic patterns. The main idea is to use Wikipedia as a readily annotated corpus by relying on its weasel tag. The experiments show that the syntactic patterns work better when using a broader notion of hedging tested on manual annotations. When evaluating on Wikipedia weasel tags itself, word frequency and distance to the tag is sufficient.

Our approach takes a much broader domain into account than previous work. It can also easily be applied to different languages as the weasel tag exists in more than 20 different language versions of Wikipedia. For a narrow domain, we suggest to start with our approach for deriving a seed set of hedging indicators and then to use a weakly supervised approach.

Though our classifiers were trained on data from multiple Wikipedia dumps, there were only a few hundred training instances available. The transient nature of the weasel tag suggests to use the Wikipedia edit history for future work, since the edits faithfully record all occurrences of weasel tags.

Acknowledgments. This work has been partially funded by the European Union under the project Judicial Management by Digital Libraries Semantics (JUMAS FP7-214306) and by the Klaus Tschira Foundation, Heidelberg, Germany.

References

Bachenko, Joan, Eileen Fitzpatrick & Michael Schonwetter (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics,* Manchester, U.K., 18–22 August 2008, pp. 41–48.

Brants, Thorsten (2000). TnT – A statistical Part-of-Speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing,* Seattle, Wash., 29 April – 4 May 2000, pp. 224–231.

Carletta, Jean (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics,* 22(2):249–254.

Hyland, Ken (1998). *Hedging in scientific research articles.* Amsterdam, The Netherlands: John Benjamins.

Lakoff, George (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2:458–508.

Light, Marc, Xin Ying Qiu & Padmini Srinivasan (2004). The language of Bioscience: Facts, speculations, and statements in between. In *Proceedings of the HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases,* Boston, Mass., 6 May 2004, pp. 17–24.

Medlock, Ben & Ted Briscoe (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, pp. 992–999.

Pang, Bo & Lillian Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Riloff, Ellen, Janyce Wiebe & Theresa Wilson (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Computational Natural Language Learning,* Edmonton, Alberta, Canada, 31 May – 1 June 2003, pp. 25–32.

Szarvas, György (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* Columbus, Ohio, 15–20 June 2008, pp. 281–289.