Statistical Modality Tagging from Rule-based Annotations and Crowdsourcing

Vinodkumar Prabhakaran
CS
Columbia University
vinod@cs.columbia.edu

Michael Bloodgood
CASL
University of Maryland
meb@umd.edu

Mona Diab
CCLS
Columbia University
mdiab@ccls.columbia.edu

Bonnie Dorr
CS and UMIACS
University of Maryland
bonnie@umiacs.umd.edu

Lori Levin
LTI
Carnegie Mellon University
lsl@cs.cmu.edu

Christine D. Piatko
APL
Johns Hopkins University
christine.piatko@jhuapl.edu

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

Benjamin Van Durme
HLTCOE
Johns Hopkins University
vandurme@cs.jhu.edu

Abstract

We explore training an automatic modality tagger. Modality is the attitude that a speaker might have toward an event or state. One of the main hurdles for training a linguistic tagger is gathering training data. This is particularly problematic for training a tagger for modality because modality triggers are sparse for the overwhelming majority of sentences. We investigate an approach to automatically training a modality tagger where we first gathered sentences based on a high-recall simple rule-based modality tagger and then provided these sentences to Mechanical Turk annotators for further annotation. We used the resulting set of training data to train a precise modality tagger using a multi-class SVM that delivers good performance.

1 Introduction

Modality is an extra-propositional component of meaning. In *John may go to NY*, the basic proposition is *John go to NY* and the word *may* indicates modality. Van Der Auwera and Ammann (2005) define core cases of modality: *John must go to NY* (epistemic necessity), *John might go to NY* (epistemic possibility), *John has to leave now* (deontic necessity) and *John may leave now* (deontic possibility). Many semanticists (e.g. Kratzer (1981), Kratzer (1991), Kaufmann et al. (2006)) define modality as quantification over possible worlds. *John might go* means that there exist some possible worlds in which John goes. Another view of modality relates more to a speakers attitude toward a proposition (e.g. McShane et al. (2004)).

Modality might be construed broadly to include several types of attitudes that a speaker wants to express towards an event, state or proposition. Modality might indicate factivity, evidentiality, or sentiment (McShane et al., 2004). Factivity is related to whether the speaker wishes to convey his or her belief that the propositional content is true or not, i.e., whether it actually obtains in this world or not. It distinguishes things that (the speaker believes) happened from things that he or she desires, plans, or considers merely probable. Evidentiality deals with the source of information and may provide clues to the reliability of the information. Did the speaker

Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012), pages 57–64, Jeju, Republic of Korea, 13 July 2012.

have firsthand knowledge of what he or she is reporting, or was it hearsay or inferred from indirect evidence? Sentiment deals with a speaker's positive or negative feelings toward an event, state, or proposition.

In this paper, we focus on the following five modalities; we have investigated the belief/factivity modality previously (Diab et al., 2009b; Prabhakaran et al., 2010), and we leave other modalities to future work.

- **Ability:** can H do P?
- **Effort:** does H try to do P?
- **Intention:** does H intend P?
- **Success:** does H succeed in P?
- **Want:** does H want P?

We investigate automatically training a modality tagger by using multi-class Support Vector Machines (SVMs). One of the main hurdles for training a linguistic tagger is gathering training data. This is particularly problematic for training a modality tagger because modality triggers are sparse for the overwhelming majority of the sentences. Baker et al. (2010) created a modality tagger by using a semi-automatic approach for creating rules for a rule-based tagger. A pilot study revealed that it can boost recall well above the naturally occurring proportion of modality without annotated data but with only 60% precision. We investigated an approach where we first gathered sentences based on a simple modality tagger and then provided these sentences to annotators for further annotation, The resulting annotated data also preserved the level of inter-annotator agreement for each example so that learning algorithms could take that into account during training. Finally, the resulting set of annotations was used for training a modality tagger using SVMs, which gave a high precision indicating the success of this approach.

Section 2 discusses related work. Section 3 discusses our procedure for gathering training data. Section 4 discusses the machine learning setup and features used to train our modality tagger and presents experiments and results. Section 5 concludes and discusses future work.

2 Related Work

Previous related work includes TimeML (Sauri et al., 2006), which involves modality annotation on events, and Factbank (Sauri and Pustejovsky, 2009), where event mentions are marked with degree of factuality. Modality is also important in the detection of uncertainty and hedging. The CoNLL shared task in 2010 (Farkas et al., 2010) deals with automatic detection of uncertainty and hedging in Wikipedia and biomedical sentences.

Baker et al. (2010) and Baker et al. (2012) analyze a set of eight modalities which include belief, require and permit, in addition to the five modalities we focus on in this paper. They built a rule-based modality tagger using a semi-automatic approach to create rules. This earlier work differs from the work described in this paper in that the our emphasis is on the creation of an *automatic* modality tagger using machine learning techniques. Note that the annotation and automatic tagging of the belief modality (i.e., factivity) is described in more detail in (Diab et al., 2009b; Prabhakaran et al., 2010).

There has been a considerable amount of interest in modality in the biomedical domain. Negation, uncertainty, and hedging are annotated in the Bioscope corpus (Vincze et al., 2008), along with information about which words are in the scope of negation/uncertainty. The i2b2 NLP Shared Task in 2010 included a track for detecting assertion status (e.g. present, absent, possible, conditional, hypothetical etc.) of medical problems in clinical records.[1] Apostolova et al. (2011) presents a rule-based system for the detection of negation and speculation scopes using the Bioscope corpus. Other studies emphasize the importance of detecting uncertainty in medical text summarization (Morante and Daelemans, 2009; Aramaki et al., 2009).

Modality has also received some attention in the context of certain applications. Earlier work describing the difficulty of correctly translating modality using machine translation includes (Sigurd and Gawrónska, 1994) and (Murata et al., 2005). Sigurd et al. (1994) write about rule based frameworks and how using alternate grammatical constructions such as the passive can improve the rendering of the modal in the target language. Murata et al. (2005)

[1] https://www.i2b2.org/NLP/Relations/

analyze the translation of Japanese into English by several systems, showing they often render the present incorrectly as the progressive. The authors trained a support vector machine to specifically handle modal constructions, while our modal annotation approach is a part of a full translation system.

The textual entailment literature includes modality annotation schemes. Identifying modalities is important to determine whether a text entails a hypothesis. Bar-Haim et al. (2007) include polarity based rules and negation and modality annotation rules. The polarity rules are based on an independent polarity lexicon (Nairn et al., 2006). The annotation rules for negation and modality of predicates are based on identifying modal verbs, as well as conditional sentences and modal adverbials. The authors read the modality off parse trees directly using simple structural rules for modifiers.

3 Constructing Modality Training Data

In this section, we will discuss the procedure we followed to construct the training data for building the automatic modality tagger. In a pilot study, we obtained and ran the modality tagger described in (Baker et al., 2010) on the English side of the Urdu-English LDC language pack.[2] We randomly selected 1997 sentences that the tagger had labeled as not having the Want modality and posted them on Amazon Mechanical Turk (MTurk). Three different Turkers (MTurk annotators) marked, for each of the sentences, whether it contained the Want modality. Using majority rules as the Turker judgment, 95 (i.e., 4.76%) of these sentences were marked as having a Want modality. We also posted 1993 sentences that the tagger had labeled as having a Want modality and only 1238 of them were marked by the Turkers as having a Want modality. Therefore, the estimated precision of this type of approach is only around 60%.

Hence, we will not be able to use the (Baker et al., 2010) tagger to gather training data. Instead, our approach was to apply a simple tagger as a first pass, with positive examples subsequently hand-annotated using MTurk. We made use of sentence data from the Enron email corpus,[3] derived from the version owing to Fiore and Heer,[4] further processed as described by (Roark, 2009).[5]

To construct the simple tagger (the first pass), we used a lexicon of modality trigger words (e.g., *try, plan, aim, wish, want*) constructed by Baker et al. (2010). The tagger essentially tags each sentence that has a word in the lexicon with the corresponding modality. We wrote a few simple obvious filters for a handful of exceptional cases that arise due to the fact that our sentences are from e-mail. For example, we filtered out *best wishes* expressions, which otherwise would have been tagged as *Want* because of the word *wishes*.

The words that trigger modality occur with very different frequencies. If one is not careful, the training data may be dominated by only the commonly occurring trigger words and the learned tagger would then be biased towards these words. In order to ensure that our training data had a diverse set of examples containing many lexical triggers and not just a lot of examples with the same lexical trigger, for each modality we capped the number of sentences from a single trigger to be at most 50. After we had the set of sentences selected by the simple tagger, we posted them on MTurk for annotation.

The Turkers were asked to check a box indicating that the modality was not present in the sentence if the given modality was not expressed. If they did not check that box, then they were asked to highlight the target of the modality. Table 1 shows the number of sentences we posted on MTurk for each modality.[6] Three Turkers annotated each sentence. We restricted the task to Turkers who were adults, had greater than a 95% approval rating, and had completed at least 50 HITs (Human Intelligence Tasks) on MTurk. We paid US$0.10 for each set of ten sentences.

Since our data was annotated by three Turkers, for training data we used only those examples for which at least two Turkers agreed on the modality and the target of the modality. This resulted in 1,008 examples. 674 examples had two Turkers agreeing and 334 had unanimous agreement. We kept track of the level of agreement for each example so that

[2] LDC Catalog No.: LDC2006E110.

[3] http://www-2.cs.cmu.edu/∼enron/

[4] http://bailando.sims.berkeley.edu/enron/enron.sql.gz

[5] Data received through personal communication

[6] More detailed statistics on MTurk annotations are available at http://hltcoe.jhu.edu/datasets/.

Modality	Count
Ability	190
Effort	1350
Intention	1320
Success	1160
Want	1390

Table 1: For each modality, the number of sentences returned by the simple tagger that we posted on MTurk.

Modality	MTurk	Gold
Ability	6%	48%
Effort	25%	10%
Intention	30%	11%
Success	24%	9%
Want	15%	23%

Table 2: Frequency of Modalities

our learner could weight the examples differently depending on the level of inter-annotator agreement.

4 Multiclass SVM for Modality

In this section, we describe the automatic modality tagger we built using the MTurk annotations described in Section 3 as the training data. Section 4.1 describes the training and evaluation data. In Section 4.2, we present the machinery and Section 4.3 describes the features we used to train the tagger. In Section 4.4, we present various experiments and discuss results. Section 4.5, presents additional experiments using annotator confidence.

4.1 Data

For training, we used the data presented in Section 3. We refer to it as MTurk data in the rest of this paper. For evaluation, we selected a part of the LU Corpus (Diab et al., 2009a) (1228 sentences) and our expert annotated it with modality tags. We first used the high-recall simple modality tagger described in Section 3 to select the sentences with modalities. Out of the 235 sentences returned by the simple modality tagger, our expert removed the ones which did not in fact have a modality. In the remaining sentences (94 sentences), our expert annotated the target predicate. We refer to this as the Gold dataset in this paper. The MTurk and Gold datasets differ in terms of genres as well as annotators (Turker vs. Expert). The distribution of modalities in both MTurk and Gold annotations are given in Table 2.

4.2 Approach

We applied a supervised learning framework using multi-class SVMs to automatically learn to tag modalities in context. For tagging, we used the Yamcha (Kudo and Matsumoto, 2003) sequence labeling system which uses the SVM$^{\text{light}}$ (Joachims, 1999) package for classification. We used *One versus All* method for multi-class classification on a quadratic kernel with a C value of 1. We report recall and precision on word tokens in our corpus for each modality. We also report $F_{\beta=1}$ (F)-measure as the harmonic mean between (P)recision and (R)ecall.

4.3 Features

We used lexical features at the token level which can be extracted without any parsing with relatively high accuracy. We use the term context width to denote the window of tokens whose features are considered for predicting the tag for a given token. For example, a context width of 2 means that the feature vector of any given token includes, in addition to its own features, those of 2 tokens before and after it as well as the tag prediction for 2 tokens before it. We did experiments varying the context width from 1 to 5 and found that a context width of 2 gives the optimal performance. All results reported in this paper are obtained with a context width of 2. For each token, we performed experiments using following lexical features:

- **wordStem** - Word stem.
- **wordLemma** - Word lemma.
- **POS** - Word's POS tag.
- **isNumeric** - Word is Numeric?
- **verbType** - Modal/Auxiliary/Regular/Nil
- **whichModal** - If the word is a modal verb, which modal?

We used the Porter stemmer (Porter, 1997) to obtain the stem of a word token. To determine the word lemma, we used an in-house lemmatizer using dictionary and morphological analysis to obtain the dictionary form of a word. We obtained POS tags from Stanford POS tagger and used those tags to determine *verbType* and *whichModal* features. The *verbType* feature is assigned a value 'Nil' if the word is not a verb and *whichModal* feature is assigned a value 'Nil' if the word is not a modal verb. The feature *isNumeric* is a binary feature denoting whether the token contains only digits or not.

4.4 Experiments and Results

In this section, we present experiments performed considering all the MTurk annotations where two annotators agreed and all the MTurk annotations where all three annotators agreed to be equally correct annotations. We present experiments applying differential weights for these annotations in Section 4.5. We performed 4-fold cross validation (4FCV) on MTurk data in order to select the best feature set configuration ϕ. The best feature set obtained was $wordStem, POS, whichModal$ with a context width of 2. For finding the best performing feature set - context width configuration, we did an exhaustive search on the feature space, pruning away features which were proven not useful by results at stages. Table 3 presents results obtained for each modality on 4-fold cross validation.

Modality	Precision	Recall	F Measure
Ability	82.4	55.5	65.5
Effort	95.1	82.8	88.5
Intention	84.3	61.3	70.7
Success	93.2	76.6	83.8
Want	88.4	64.3	74.3
Overall	**90.1**	**70.6**	**79.1**

Table 3: Per modality results for best feature set ϕ on 4-fold cross validation on MTurk data

We also trained a model on the entire MTurk data using the best feature set ϕ and evaluated it against the Gold data. The results obtained for each modality on gold evaluation are given in Table 4. We attribute the lower performance on the Gold dataset to its difference from MTurk data. MTurk data is entirely from email threads, whereas Gold data contained sentences from newswire, letters and blogs in addition to emails. Furthermore, the annotation is different (Turkers vs expert). Finally, the distribution of modalities in both datasets is very different. For example, *Ability* modality was merely 6% of MTurk data compared to 48% in Gold data (see Table 2).

Modality	Precision	Recall	F Measure
Ability	78.6	22.0	34.4
Effort	85.7	60.0	70.6
Intention	66.7	16.7	26.7
Success	NA	0.0	NA
Want	92.3	50.0	64.9
Overall	**72.1**	**29.5**	**41.9**

Table 4: Per modality results for best feature set ϕ evaluated on Gold dataset

We obtained reasonable performances for *Effort* and *Want* modalities while the performance for other modalities was rather low. Also, the Gold dataset contained only 8 instances of *Success*, none of which was recognized by the tagger resulting in a recall of 0%. Precision (and, accordingly, F Measure) for *Success* was considered "not applicable" (NA), as no such tag was assigned.

4.5 Annotation Confidence Experiments

Our MTurk data contains sentence for which at least two of the three Turkers agreed on the modality and the target of the modality. In this section, we investigate the role of annotation confidence in training an automatic tagger. The annotation confidence is denoted by whether an annotation was agreed by only two annotators or was unanimous. We denote the set of sentences for which only two annotators agreed as Agr_2 and that for which all three annotators agreed as Agr_3.

We present four training setups. The first setup is $Tr23$ where we train a model using both Agr_2 and Agr_3 with equal weights. This is the setup we used for results presented in the Section 4.4. Then, we have $Tr2$ and $Tr3$, where we train using only Agr_2 and Agr_3 respectively. Then, for $Tr23_W$, we

TrainingSetup	Tested on Agr_2 and Agr_3			Tested on Agr_3 only		
	Precision	Recall	F Measure	Precision	Recall	F Measure
$Tr23$	90.1	**70.6**	**79.1**	95.9	**86.8**	**91.1**
$Tr2$	**91.0**	66.1	76.5	95.6	81.8	88.2
$Tr3$	88.1	52.3	65.6	**96.8**	71.7	82.3
$Tr23_W$	89.9	70.5	79.0	95.8	86.5	90.9

Table 5: Annotator Confidence Experiment Results; the best results per column are boldfaced
(4-fold cross validation on MTurk Data)

train a model giving different cost values for Agr_2 and Agr_3 examples. The SVMLight package allows users to input cost values c_i for each training instance separately.[7] We tuned this cost value for Agr_2 and Agr_3 examples and found the best value at 20 and 30 respectively.

For all four setups, we used feature set ϕ. We performed 4-fold cross validation on MTurk data in two ways — we tested against a combination of Agr_2 and Agr_3, and we tested against only Agr_3. Results of these experiments are presented in Table 5. We also present the results of evaluating a tagger trained on the whole MTurk data for each setup against the Gold annotation in Table 6. The $Tr23$ tested on both Agr_2 and Agr_3 presented in Table 5 and $Tr23$ tested on Gold data presented in Table 6 correspond to the results presented in Table 3 and Table 4 respectively.

TrainingSetup	Precision	Recall	F Measure
$Tr23$	72.1	29.5	41.9
$Tr2$	67.4	27.6	39.2
$Tr3$	**74.1**	19.1	30.3
$Tr23_W$	73.3	**31.4**	**44.0**

Table 6: Annotator Confidence Experiment Results; the best results per column are boldfaced
(Evaluation against Gold)

One main observation is that including annotations of lower agreement, but still above a threshold (in our case, 66.7%), is definitely helpful. $Tr23$ outperformed both $Tr2$ and $Tr3$ in both recall and F-measure in all evaluations. Also, even when evaluating against only the high confident Agr_3 cases, $Tr2$ gave a high gain in recall (10 .1 percentage points) over $Tr3$, with only a 1.2 percentage point loss on precision. We conjecture that this is because there are far more training instances in $Tr2$ than in $Tr3$ (674 vs 334), and that quantity beats quality.

Another important observation is the increase in performance by using varied costs for Agr_2 and Agr_3 examples (the $Tr23_W$ condition). Although it dropped the performance by 0.1 to 0.2 points in cross-validation F measure on the Enron corpora, it gained 2.1 points in Gold evaluation F measure. These results seem to indicate that differential weighting based on annotator agreement might have more beneficial impact when training a model that will be applied to a wide range of genres than when training a model with genre-specific data for application to data from the same genre. Put differently, using varied costs prevents genre over-fitting. We don't have a full explanation for this difference in behavior yet. We plan to explore this in future work.

5 Conclusion

We have presented an innovative way of combining a high-recall simple tagger with Mechanical Turk annotations to produce training data for a modality tagger. We show that we obtain good performance on the same genre as this training corpus (annotated in the same manner), and reasonable performance across genres (annotated by an independent expert). We also present experiments utilizing the number of agreeing Turkers to choose cost values for training examples for the SVM. As future work, we plan to extend this approach to other modalities which are

[7] This can be done by specifying 'cost:<value>' after the label in each training instance. This feature has not yet been documented on the SVMlight website.

not covered in this study.

6 Acknowledgments

This work is supported, in part, by the Johns Hopkins Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor. We thank several anonymous reviewers for their constructive feedback.

References

Emilia Apostolova, Noriko Tomuro, and Dina Demner-Fushman. 2011. Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 283–287, Portland, Oregon.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192, Boulder, Colorado, June. Association for Computational Linguistics.

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Nathaniel W. Filardo, Lori S. Levin, and Christine D. Piatko. 2010. A modality lexicon and its use in automatic tagging. In *LREC*.

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Use of modality and negation in semantically-informed syntactic mt. *Computational Linguistics*, 38(22).

Roy Bar-Haim, Ido Dagan, Iddo Greental, and Eyal Shnarch. 2007. Semantic inference at the lexical-syntactic level. In *Proceedings of the 22nd National Conference on Artificial intelligence - Volume 1*, pages 871–876, Vancouver, British Columbia, Canada. AAAI Press.

Mona Diab, Bonnie Dorr, Lori Levin, Teruko Mitamura, Rebecca Passonneau, Owen Rambow, and Lance Ramshaw. 2009a. *Language Understanding Annotation Corpus*. Linguistic Data Consortium (LDC), USA.

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

Richárd Farkas, Veronika Vincze, György Szarvas, György Móra, and János Csirik, editors. 2010. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Uppsala, Sweden, July.

Thorsten Joachims, 1999. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, Cambridge, MA, USA.

Stefan Kaufmann, Cleo Condoravdi, and Valentina Harizanov, 2006. *Formal Approaches to Modality*, pages 72–106. Mouton de Gruyter.

Angelika Kratzer. 1981. The Notional Category of Modality. In H. J. Eikmeyer and H. Rieser, editors, *Words, Worlds, and Contexts*, pages 38–74. de Gruyter, Berlin.

Angelika Kratzer. 1991. Modality. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*. de Gruyter.

Taku Kudo and Yuji Matsumoto. 2003. Fast methods for kernel-based text analysis. In *41st Meeting of the Association for Computational Linguistics (ACL'03)*, Sapporo, Japan.

Marjorie McShane, Sergei Nirenburg, and Ron Zacharsky. 2004. Mood and modality: Out of the theory and into the fray. *Natural Language Engineering*, 19(1):57–89.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP 2009 Workshop*, pages 28–36, Boulder, Colorado, June. Association for Computational Linguistics.

Masaki Murata, Kiyotaka Uchimoto, Qing Ma, Toshiyuki Kanamaru, and Hitoshi Isahara. 2005. Analysis of machine translation systems' errors in tense, aspect, and modality. In *Proceedings of the 19th Asia-Pacific Conference on Language, Information and Computation (PACLIC)*, Tapei.

Rowan Nairn, Cleo Condorovdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of the International Workshop on Inference in Computational Semantics*, ICoS-5, pages 66–76, Buxton, England.

M. F. Porter, 1997. *An algorithm for suffix stripping*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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.

Brian Roark. 2009. Open vocabulary language modeling for binary response typing interfaces. Technical report, Oregon Health and Science University.

Roser Sauri and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Roser Sauri, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *FLAIRS Conference*, pages 333–339.

Bengt Sigurd and Barbara Gawrónska. 1994. Modals as a problem for MT. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING) Volume 1*, COLING '94, pages 120–124, Kyoto, Japan.

Johan Van Der Auwera and Andreas Ammann, 2005. *Overlap between situational and epistemic modal marking*, chapter 76, pages 310–313. Oxford University Press.

Veronika Vincze, Gy orgy Szarvas, Richád Farkas, Gy orgy Mora, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9+.