Enriching the Bioscope Corpus with Lexical and Syntactic Information

Guillermo Moncecchi[1,2], Jean-Luc Minel[2], and Dina Wonsever[1]

[1] Instituto de Computación, Facultad de Ingeniería
Universidad de la República
Montevideo
Uruguay
[2] Laboratoire MoDyCo, UMR 7114 CNRS
Université Paris Ouest
Nanterre La Défense
France

Abstract. This paper details the method used to augment an epistemic modality corpus (the Bioscope corpus), incorporating results from the lexical and syntactic analysis of its sentences. The features resulting from these analyses were consolidated in a single data structure, that can be used for interactive experimentation on the corpus. Some visualization aids developed for corpus browsing are also described.

1 Introduction

This work is part of a more general study of epistemic modality in unrestricted texts, particularly in the molecular biology domain. Previous work on the Microbio project[1], a collaboration between biologists, computer scientists and linguists, highlighted the need to identify different types of features (temporal, modal, declarative) associated with information extracted from molecular biology papers. In particular, the possibility of detecting epistemic modality markers and connecting them to certain textual segments could be extremely useful in the transition from unstructured text to a knowledge base.

We consider this work as the first step towards the study of the influence of syntactic analysis information on detecting hedge cues and their scope. Our hypothesis is that working out the constituent structure of sentences could help to identify dependencies between hedge cues and their scope, which could be subsequently used as features for learning. For example in the sentence *"IFN-alpha also sensitized T cells to IL-2-induced proliferation, further suggesting that IFN-alpha may be involved in the regulation of T-cell mitogenesis"*, the scope of the cue *may* is the sentence *"IFN-alpha may be involved in the regulation of T-cell mitogenesis."*, syntactically constituted by the noun *IFN-alpha*, and the verb phrase *"may be involved in the regulation of T-cell mitogenesis."*. Based on this observation, a linguist could hypothesize that when the term *may* appears in a proposition as a modifier and is marked as a hedge cue, its hedging scope should be the sentence that includes it, and further, that relations extracted from the

propositions (in the example, the relation between the protein *IFN-alpha* and the regulation process), should be tagged as uncertain.

As far as we know, no previous work has been done in this direction: only lexical features such as words, lemmas and POS tags have been used when trying to classify modality aspects using machine learning methods. We think that if information from different analysis sources (part-of-speech tagging, syntactic analysis, hand-tagged hedge scope, named entity recognition, and chunking) could be integrated, so as to build an integrated corpus and analysis environment, this would facilitate the analysis of hedging phenomena, and thus improve hedge detection using standard machine learning methods.

In this work we show how we enriched the Bioscope corpus (a biological corpus annotated with modality cues and their scope), by integrating syntactic and lexical information resulting from different analysis tools, and building an environment for its visualization and browsing. The next section briefly reviews the concepts of epistemic modality and hedging, and summarizes previous work on hedge classification, particularly in the biological domain. Then we describe our work and the lines of investigation planned for the future.

2 Related Work

Palmer[7] defines *epistemic modality* as *"any modal system that indicates the degree of commitment by the speaker to what he says (...); this clearly includes both his own judgements and the kind of warrant he has for what he says"*. Related to the concept of epistemic modality is the notion of *hedge*. The term was first introduced by Lakoff[3],who studied the properties of words like *rather* and their ability to *"make things fuzzier or less fuzzy"*. The Concise Oxford Dictionary of Linguistics (cited by [8]) defines hedges as: *"any linguistic device by which a speaker avoids being compromised by a statement that turns out to be wrong, a request that is not acceptable, and so on."*. It should be clear that hedges are strong indicators about the epistemic modality of any assertion.

Saurí et al.[10] investigated the general modality of events, *"which expresses the speaker's degree of commitment to the events being referred to in a text"*, and defined different modal types, including degrees of possibility, belief, evidentiality, expectation, attempting and command. They remarked that modality identification should be a layer of information in text analysis, to allow better inferences from events.

This level of analysis seems very important when considering scientific writing: scientific assertions often include some degree of uncertainty or assessment of possibilities[8]. Detecting epistemic modality features from identified assertions could help with concept identification and relation extraction. The expression *"Here we show that the response of the HIV-1 LTR may be governed by two independent sequences located 5' to the site of transcription initiation sequences that bind either NFAT-1 or NF kappa B"* asserts a relation between *"the response of HIV-1 LTR"* and two DNA sequences. An information extraction system that skipped the modality analysis would miss the fact that the author includes the

relation under the scope of a hedge (in this case, *may*), showing he is not sure about it, and so it should be presented with lower confidence.

In recent years, hedging has been the target of several studies, even deserving a shared task in the Tenth Conference on Computational Natural Language Learning (CoNLL-2010). Medlock and Briscoe [5] showed that hedge classification could be seen as a weakly supervised machine learning task. Using Support Vector Machines, they achieved a recall/precision break even point of 0.76 on a corpus they built and made publicly available, using a bag-of-words model as features. Later work by Medlock [4] added POS tags, lemmas and bigrams as learning features, achieving a maximum BEP of 0.82. Working on the same corpus, Szarvas [11] achieved a of 0.85 using an external dictionary of hedge keywords, a Maximum Entropy Markov Model classifier on trigrams, bigrams and unigrams, and a weighting mechanism on hedge cues. Morante and Daelemans [6] not only tried to detect hedge cues but also their scope, learning on the Bioscope corpus (see below). They used a metalearning approach based on three supervised learning methods: memory-based learning, Support Vector Machines and Conditional Random Fields. The features used were chain-of words, lemmas, POS tags, chunk IOB tags, token location relative to the hedge cue, and a list of cue candidate words. They achieved an F1 of 74.05 for hedge identification, and 90.61 for scope finding (using gold-standard hedge signals).

As mentioned above, we are not aware of research that has used parsing information as features for hedge detection and scope finding, nor do we know of the existence of an epistemic modality corpus annotated with this type of information.

3 Enriching a Modality Corpus

3.1 The Bioscope Corpus

The Bioscope corpus [13] is a freely available corpus of medical free texts, biological full papers and biological scientific abstracts, annotated at a token level with negative and speculative keywords, and at sentence level with their linguistic scope. It includes 20.000 sentences considered for annotation, 10% of them actually containing one or more linguistic annotations suggesting negation or uncertainty.

Related to hedge detection, uncertainty markers and their scopes were identified. Negation and uncertainty scopes can be nested, yielding results such as the following annotated sentence:

```
<sentence>The induction of AP1 by okadaic acid
<xcope><cue type="speculation">suggests</cue>
that protein phosphatases 1 and 2A (PP1 and PP2A)
<xcope><cue type="speculation">may</cue> be involved in T cell
activation as important negative regulators of the transcription
factor AP1</xcope></xcope>.</sentence>
```

Table 3.1, extracted from [13], gives some statistics related to hedge cues and sentences for the three sub corpora included in Bioscope.

	Clinical	Full paper	Abstract
#Documents	954	9	1273
#Sentences	6383	2670	11871
Hedge Sentences	13.39%	19.44%	17.70%
#Hedge cues	1189	714	2769

Table 1. Bioscope corpus statistics about hedging

3.2 Added Information

As previously mentioned, we aimed at enriching the texts in the Bioscope corpus with results from different analyses in order to obtain a new richer corpus, suitable for use on hedge detection tasks. We started with the original sentences of the corpus, tokenized them and added lexical, syntactic and hedging information.

Lexical information To incorporate lexical information, each Bioscope sentence was analysed with the GENIA tagger [12], a widely used part-of-speech tagger, especially trained on the biological domain.This tagger was also used to annotate named entities and chunking information at a token level.

Hedge information Hedge information (already present in the corpus) cannot be directly represented at a token level: it has an arborescent structure, with potentially nested scopes. However, browsing the corpus, we found that (as could be expected) hedging scopes nesting was almost never deeper than two levels. We therefore modelled hedging information with two scope attributes, allowing at most one nested hedging scope. To tag speculative cues and scopes a standard model used in Named Entity Recognition tasks was applied: as hedge cues could extend for more than just a word, the first token of a hedge cue was tagged with the tag B-SPECCUE, and the rest of the words in the cue with I-SPECCUE . Tokens not included in a cue where given the tag O. A similar approach was used for scope identification.

Table 3.2 shows the aforementioned sentence, including its hedging and lexical attributes. As can be seen, the information included so far, represented as a list of tokens with their attributes, following the standard of the 2006 CoNLL Shared Task, can be used straightforwardly in machine learning tasks such as classification.

Sentence constituents We also analysed the corpus searching for sentence constituents, using the Sanford Parser[2]. We built a syntactic analysis tree for

each sentence of the corpus. As this parser was trained on a different domain from ours, we tried to improve its performance using as inputs for the parser the tokens that resulted from the GENIA tagger analysis, and their part-of-speech tags. As the usual representation of token-per-token features did not satisfactorily accommodate the parsing information (which is essentially tree-shaped), we decided to start with the tree resulting from the syntactic analysis, then decorating each of its leaves (containing sentence tokens), with the rest of the features.

Figure 1 shows a small part of its analysis tree decorated with the part-of-speech, chunk, NER and hedging features.

Token	Lemma	POS	Chunk	NE	Hedge Cue 1	Scope	Hedge Cue 2	Scope
The	The	DT	B-NP	O	O	O	O	O
induction	induction	NN	I-NP	O	O	O	O	O
of	of	IN	B-PP	O	O	O	O	O
AP1	AP1	NN	B-NP	B-protein	O	O	O	O
by	by	IN	B-PP	O	O	O	O	O
okadaic	okadaic	JJ	B-NP	O	O	O	O	O
acid	acid	NN	I-NP	O	O	O	O	O
suggests	suggest	VBZ	B-VP	O	B-SPECCUE	B-XCOPE	O	O
that	that	IN	B-SBAR	O	O	I-XCOPE	O	O
protein	protein	NN	B-NP	B-protein	O	I-XCOPE	O	O
phosphatases	phosphatas	NNS	I-NP	I-protein	O	I-XCOPE	O	O
1	1	CD	B-NP	I-protein	O	I-XCOPE	O	O
and	and	CC	O	I-protein	O	I-XCOPE	O	O
2A	2A	NN	B-NP	I-protein	O	I-XCOPE	O	O
-LRB-	(	-LRB-	O	O	O	I-XCOPE	O	O
PP1	PP1	NN	B-NP	B-protein	O	I-XCOPE	O	O
and	and	CC	O	O	O	I-XCOPE	O	O
PP2A	PP2A	NN	B-NP	B-protein	O	I-XCOPE	O	O
-RRB-	)	-RRB-	O	O	O	I-XCOPE	O	O
may	may	MD	B-VP	O	O	I-XCOPE	B_SPECCUE	B-XCOPE
be	be	VB	I-VP	O	O	I-XCOPE	O	I-XCOPE
involved	involve	VBN	I-VP	O	O	I-XCOPE	O	I-XCOPE
in	in	IN	B-PP	O	O	I-XCOPE	O	I-XCOPE
T	T	NN	B-NP	O	O	I-XCOPE	O	I-XCOPE
cell	cell	NN	I-NP	O	O	I-XCOPE	O	I-XCOPE
activation	activation	NN	I-NP	O	O	I-XCOPE	O	I-XCOPE
as	as	IN	B-PP	O	O	I-XCOPE	O	I-XCOPE
important	important	JJ	B-NP	O	O	I-XCOPE	O	I-XCOPE
negative	negative	JJ	I-NP	O	O	I-XCOPE	O	I-XCOPE
regulators	regulator	NNS	I-NP	O	O	I-XCOPE	O	I-XCOPE
of	of	IN	B-PP	O	O	I-XCOPE	O	I-XCOPE
the	the	DT	B-NP	O	O	I-XCOPE	O	I-XCOPE
transcription	transcription	NN	I-NP	B-protein	O	I-XCOPE	O	I-XCOPE
factor	factor	NN	I-NP	I-protein	O	I-XCOPE	O	I-XCOPE
AP1	AP1	NN	I-NP	O	O	I-XCOPE	O	I-XCOPE
.	.	.	O	O	O	O	O	O

Table 2. Lexical and hedging information

The main issue in synchronizing the three sources of information was tokenization and the selection of the tagset: if we could not manage to tokenise the sentences in exactly the same way by the different analyses, integrating them into one structure would not be possible. Fortunately, the tagger and parser used the same tag-set (the PennTreebank) and its conventions for tokenization, so we followed the same approach when tokenising the Bioscope sentences. Even then, certain problems arose with GENIA incorrectly tokenising some sentences, or not following exactly the tokenization conventions. The GENIA results were post-processed to correct these situations.

As an automatic process, the addition of new information to the corpus comes at the cost of introducing analysis errors. In this work, errors come from two sources: tagging and syntactic analysis. Studying (and solving) errors that

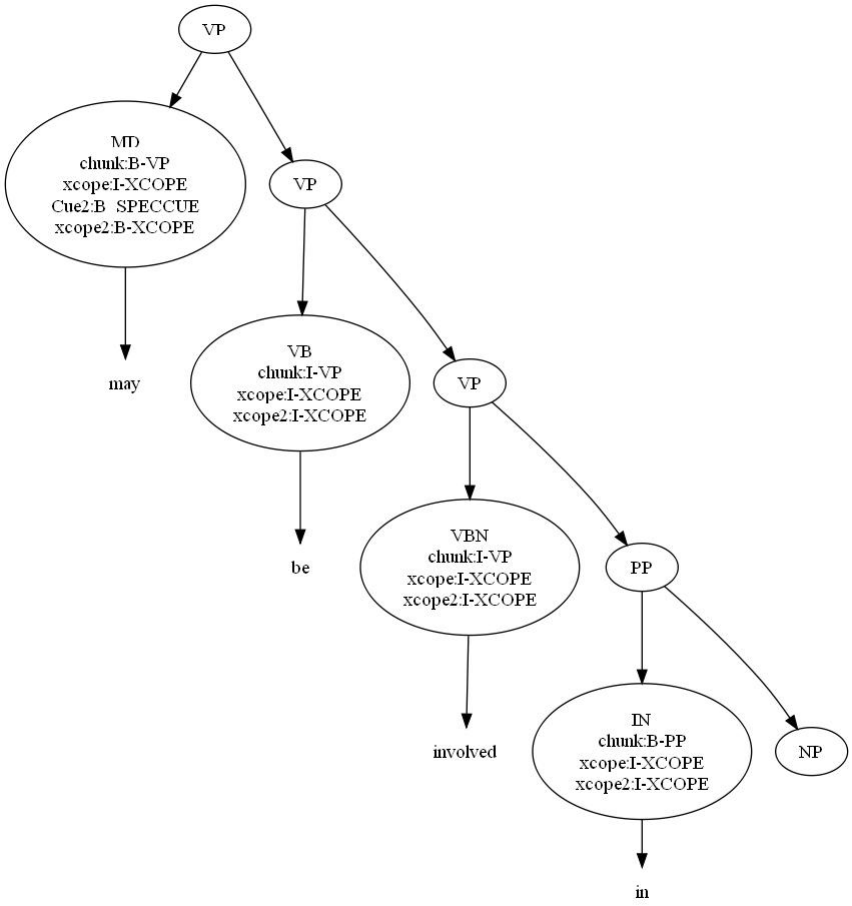

Fig. 1. Parsing information augmented with lexical and hedging features

were introduced during the process is a pending and cost-consuming task (which should include the work of linguists and domain specialists). To minimize these errors, two decisions were taken: using a domain-specific tagger (the GENIA tagger, trained on the same GENIA corpus the Bioscope corpus is partly built on), and passing this tagging information to the parser. The GENIA tagger has a reported accuracy of 0.96-0.98 on the domain [9,12], and the Stanford parser presents a F-measure of 0.86 [2] (using their own tagging method). Based on this information, we think the tagging and parsing errors introduced still allow for the use of the tagged data to improve performance on supervised learning tasks.

4 An Environment for Corpus Browsing and Visualization

Besides adding information to the original corpus, we tried to provide mechanisms for experts to easily browse the corpus content (including sentences and associated features). When doing this, two distinct kind of users were considered: linguists analysing hedging structures and their relation with lexical and syntactic features, and computer scientists trying to automatically recognize those structures.

For the first kind of users, we added visualization aids to the corpus: based on the original corpus XML file, and using XSLT and CSS templates, the user can browse the corpus, in which hedge and negations cues are highlighted, and their scopes highlighted. Figure 2 shows an example of these visualization aids. A tree visualization of the final structure was also included (built using the Graphviz[3] package), as was a token-per-token visualization of the lexical and hedging features (similar to the one showed in table 3.2). Preliminary user experiments showed that this approach is indeed effective for easier corpus analysis.

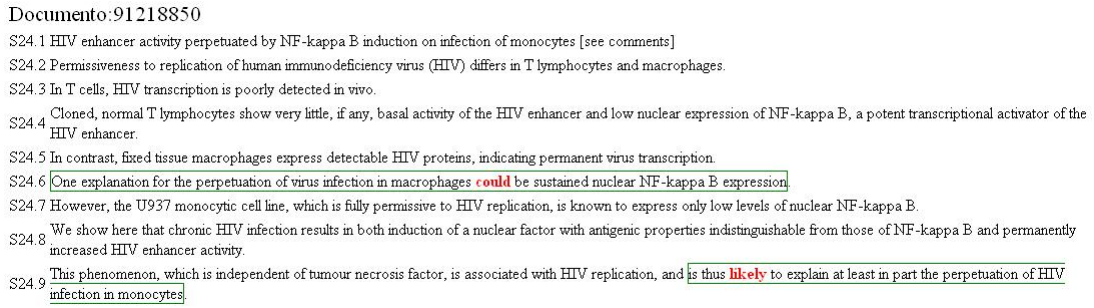

Fig. 2. Corpus visualization aids

To allow computer scientists to test their methods and techniques on the corpus, we selected the Python/NLTK environment for its modelling and im-

[3] http://www.graphviz.org

plementation. The Python language presents several advantages: its dynamic typing mechanisms allowed us to accommodate new information into the tree; our aim is to use the corpus for interactive experimentation, an interpreted language seemed a good choice; its serialization possibilities allowed us to easily save and restore the full tree structure of the corpus sentences; finally, the NLTK toolkit provided standard data structures for NLP tasks, and mechanisms for their manipulation.

5 Conclusions and Future Work

As a result of this work, we have built an augmented version of the complete Bioscope corpus, and an environment for its visualization and manipulation[4]. Data integration problems (tokenization differences, tagging discrepancies, data representation alternatives) have been completely solved. Even when the corpus is the result of automatic analysis (and, for this reason, not error-free), we claim that it can serve as a testbed for different analyses related to hedging. We hope this environment will allow for a flexible and easy way to extract statistics and rapidly evaluate the impact of different information extraction methods on the corpus. Having ways to easily visualize data and its structure will also facilitate the work of human experts on corpus analysis.

The ideas, methods and tools presented in this paper can be generalized. Incorporating information from different sources is a common task in Natural Language Processing. This integration not only assembles more features for learning methods, but can also help to improve some of the tasks involved: we noted, for example, that including GENIA-generated tags as an input for parsing seemed to improve the parsing task, but also that the parser corrected some "impossible" tags, based on its own learned statistics. Representing corpus structure as annotated trees is of course not new (compilers have used this representation for years), but its use instead of sequential representations for relation extraction and machine learning is not a common practice.

The next steps will be to model hedge cue identification and scope determination as a classification task on text tokens. To achieve this, we plan to build hand-crafted functional rules that, given a sentence, return appropriate tags for each token, based on features of the token and its neighbours in the sentence analysis tree. Using the environment, we will test each rule for precision/recall measures to assess its performance on the whole corpus. Coding rules as functions allows maximum flexibility, while having a predefined environment facilities the task of rule evaluation.

After this phase, we plan to include the results of these rules as new token features. Working on this new data, we will apply machine learning methods to generalize them. The kind of task and the size of the corpus suggest a sequential discriminative method such as Conditional Random Fields may be useful, but other approaches (Support Vector Machines, semi-supervised methods) should

[4] A reduced version of the corpus can be browsed online at http://www.fing.edu.uy/inco/grupos/pln/bioscope_devel/abstracts_devel.xml

also be considered. Finally, we intend combine these hedge classification tasks with relation extraction methods to add confidence information to extracted information.

6 Acknowledgments

This work is part of the Microbio project, partially funded by the Stic-Amsud collaborative research program, and the Temantex project, funded by the CSIC (Uruguay).

References

1. Battistelli, D., Amardeilh, F.: Knowledge Claims in Scientific Literature, Uncertainty and Semantic Annotation: A Case Study in the Biological Domain. In: Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM 2009). Los Angeles États-Unis (09 2009), http://hal.archives-ouvertes.fr/hal-00411230/en/, Microbio Stic-Amsud
2. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. pp. 423–430. Association for Computational Linguistics, Morristown, NJ, USA (2003)
3. Lakoff, G.: Hedges: A study in meaning criteria and the logic of fuzzy concepts. Journal of Philosophical Logic 2(4), 458–508 (October 1973), http://dx.doi.org/10.1007/BF00262952
4. Medlock, B.: Exploring hedge identification in biomedical literature. Journal of biomedical informatics 41(4), 636–654 (August 2008), http://dx.doi.org/10.1016/j.jbi.2008.01.001
5. Medlock, B., Briscoe, T.: Weakly supervised learning for hedge classification in scientific literature. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (2007)
6. Morante, R., Daelemans, W.: Learning the scope of hedge cues in biomedical texts. In: Proceedings of the BioNLP 2009 Workshop. pp. 28–36. Association for Computational Linguistics, Boulder, Colorado (June 2009), http://www.aclweb.org/anthology-new/W/W09/W09-1304.bib
7. Palmer, R.F.: Mood and Modality. Cambridge Textbooks in Linguistics, Cambridge University Press, New York (2001)
8. Panocov'a, R.: Expression of modality in biomedical texts. SKASE Journal of Translation and Interpretation 3, 81–90 (2009)
9. Pyysalo, S., Salakoski, T., Aubin, S., Nazarenko, A.: Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. BMC Bioinformatics 7(Suppl 3), S2+ (2006), http://dx.doi.org/10.1186/1471-2105-7-S3-S2
10. Sauri, R., Verhagen, M., Pustejovsky, J.: Slinket: A partial modal parser for events. In: Language Resources and Evaluation Conference, LREC (2006)
11. Szarvas, G.: Hedge classification in biomedical texts with a weakly supervised selection of keywords. In: Proceedings of ACL-08: HLT. pp. 281–289. Association for Computational Linguistics, Columbus, Ohio (June 2008), http://www.aclweb.org/anthology/P/P08/P08-1033
12. Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., ichi Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. In: Bozanis, P., Houstis, E.N. (eds.) Panhellenic Conference on Informatics. Lecture Notes in Computer Science, vol. 3746, pp. 382–392. Springer (2005)
13. Vincze, V., Szarvas, G., Farkas, R., Mora, G., Csirik, J.: The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9(Suppl 11), S9+ (2008), http://dx.doi.org/10.1186/1471-2105-9-S11-S9