The CoNLL 2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text

Richárd Farkas[1,2], **Veronika Vincze**[1], **György Móra**[1], **János Csirik**[1,2], **György Szarvas**[3]
[1] University of Szeged, Department of Informatics
[2] Hungarian Academy of Sciences, Research Group on Artificial Intelligence
[3] Technische Universität Darmstadt, Ubiquitous Knowledge Processing Lab
{rfarkas,vinczev,gymora,csirik}@inf.u-szeged.hu,
szarvas@tk.informatik.tu-darmstadt.de

Abstract

The CoNLL 2010 Shared Task was dedicated to the detection of uncertainty cues and their linguistic scope in natural language texts. The motivation behind this task was that distinguishing factual and uncertain information in texts is of essential importance in information extraction. This paper provides a general overview of the shared task, including the annotation protocols of the training and evaluation datasets, the exact task definitions, the evaluation metrics employed and the overall results. The paper concludes with an analysis of the prominent approaches and an overview of the systems submitted to the shared task.

1 Introduction

Every year since 1999, the Conference on Computational Natural Language Learning (CoNLL) provides a competitive shared task for the Computational Linguistics community. After a five-year period of multi-language semantic role labeling and syntactic dependency parsing tasks, a new task was introduced in 2010, namely the detection of uncertainty and its linguistic scope in natural language sentences.

In natural language processing (NLP) – and in particular, in information extraction (IE) – many applications seek to extract factual information from text. In order to distinguish facts from unreliable or uncertain information, linguistic devices such as hedges (indicating that authors do not or cannot back up their opinions/statements with facts) have to be identified. Applications should handle detected speculative parts in a different manner. A typical example is protein-protein interaction extraction from biological texts, where the aim is to mine text evidence for biological entities that are in a particular relation with each other. Here, while an uncertain relation might be of some interest for an end-user as well, such information must not be confused with factual textual evidence (reliable information).

Uncertainty detection has two levels. Automatic hedge detectors might attempt to identify sentences which contain uncertain information and handle whole sentences in a different manner or they might attempt to recognize in-sentence spans which are speculative. In-sentence uncertainty detection is a more complicated task compared to the sentence-level one, but it has benefits for NLP applications as there may be spans containing useful factual information in a sentence that otherwise contains uncertain parts. For example, in the following sentence the subordinated clause starting with *although* contains factual information while uncertain information is included in the main clause and the embedded question.

> Although IL-1 has been reported to contribute to Th17 differentiation in mouse and man, it remains to be determined {**whether** therapeutic targeting of IL-1 will substantially affect IL-17 in RA}.

Both tasks were addressed in the CoNLL 2010 Shared Task, in order to provide uniform manually annotated benchmark datasets for both and to compare their difficulties and state-of-the-art solutions for them. The uncertainty detection problem consists of two stages. First, keywords/cues indicating uncertainty should be recognized then either a sentence-level decision is made or the linguistic scope of the cue words has to be identified. The latter task falls within the scope of semantic analysis of sentences exploiting syntactic patterns, as hedge spans can usually be determined on the basis of syntactic patterns dependent on the keyword.

Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–12,
Uppsala, Sweden, 15-16 July 2010.

2 Related Work

The term *hedging* was originally introduced by Lakoff (1972). However, hedge detection has received considerable interest just recently in the NLP community. Light et al. (2004) used a handcrafted list of hedge cues to identify speculative sentences in MEDLINE abstracts and several biomedical NLP applications incorporate rules for identifying the certainty of extracted information (Friedman et al., 1994; Chapman et al., 2007; Aramaki et al., 2009; Conway et al., 2009).

The most recent approaches to uncertainty detection exploit machine learning models that utilize manually labeled corpora. Medlock and Briscoe (2007) used single words as input features in order to classify sentences from biological articles (FlyBase) as speculative or non-speculative based on semi-automatically collected training examples. Szarvas (2008) extended the methodology of Medlock and Briscoe (2007) to use n-gram features and a semi-supervised selection of the keyword features. Kilicoglu and Bergler (2008) proposed a linguistically motivated approach based on syntactic information to semi-automatically refine a list of hedge cues. Ganter and Strube (2009) proposed an approach for the automatic detection of sentences containing uncertainty based on Wikipedia weasel tags and syntactic patterns.

The BioScope corpus (Vincze et al., 2008) is manually annotated with negation and speculation cues and their linguistic scope. It consists of clinical free-texts, biological texts from full papers and scientific abstracts. Using BioScope for training and evaluation, Morante and Daelemans (2009) developed a scope detector following a supervised sequence labeling approach while Özgür and Radev (2009) developed a rule-based system that exploits syntactic patterns.

Several related works have also been published within the framework of The BioNLP'09 Shared Task on Event Extraction (Kim et al., 2009), where a separate subtask was dedicated to predicting whether the recognized biological events are under negation or speculation, based on the GENIA event corpus annotations (Kilicoglu and Bergler, 2009; Van Landeghem et al., 2009).

3 Uncertainty Annotation Guidelines

The shared task addressed the detection of uncertainty in two domains. As uncertainty detection is extremely important for biomedical information extraction and most existing approaches have targeted such applications, participants were asked to develop systems for hedge detection in biological scientific articles. Uncertainty detection is also important, e.g. in encyclopedias, where the goal is to collect reliable world knowledge about real-world concepts and topics. For example, Wikipedia explicitly declares that statements reflecting author opinions or those not backed up by facts (e.g. references) should be avoided (see 3.2 for details). Thus, the community-edited encyclopedia, Wikipedia became one of the subjects of the shared task as well.

3.1 Hedges in Biological Scientific Articles

In the biomedical domain, sentences were manually annotated for both hedge cues and their linguistic scope. Hedging is typically expressed by using specific linguistic devices (which we refer to as cues in this article) that modify the meaning or reflect the author's attitude towards the content of the text. Typical hedge cues fall into the following categories:

- auxiliaries: *may, might, can, would, should, could*, etc.
- verbs of hedging or verbs with speculative content: *suggest, question, presume, suspect, indicate, suppose, seem, appear, favor*, etc.
- adjectives or adverbs: *probable, likely, possible, unsure*, etc.
- conjunctions: *or, and/or, either ... or*, etc.

However, there are some cases where a hedge is expressed via a phrase rather than a single word. Complex keywords are phrases that express uncertainty together, but not on their own (either the semantic interpretation or the hedging strength of its subcomponents are significantly different from those of the whole phrase). An instance of a complex keyword can be seen in the following sentence:

> Mild bladder wall thickening {**raises the question of** cystitis}.

The expression *raises the question of* may be substituted by *suggests* and neither the verb *raises* nor the noun *question* convey speculative meaning on their own. However, the whole phrase is speculative therefore it is marked as a hedge cue.

During the annotation process, a min-max strategy for the marking of keywords (min) and their scope (max) was followed. On the one hand, when marking the keywords, the minimal unit that expresses hedging and determines the actual strength of hedging was marked as a keyword. On the other hand, when marking the scopes of speculative keywords, the scope was extended to the largest syntactic unit possible. That is, all constituents that fell within the uncertain interpretation were included in the scope. Our motivation here was that in this way, if we simply disregard the marked text span, the rest of the sentence can usually be used for extracting factual information (if there is any). For instance, in the example above, we can be sure that the symptom *mild bladder wall thickening* is exhibited by the patient but a diagnosis of *cystitis* would be questionable.

The scope of a speculative element can be determined on the basis of syntax. The scopes of the BioScope corpus are regarded as consecutive text spans and their annotation was based on constituency grammar. The scope of verbs, auxiliaries, adjectives and adverbs usually starts right with the keyword. In the case of verbal elements, i.e. verbs and auxiliaries, it ends at the end of the clause or sentence, thus all complements and adjuncts are included. The scope of attributive adjectives generally extends to the following noun phrase, whereas the scope of predicative adjectives includes the whole sentence. Sentential adverbs have a scope over the entire sentence, while the scope of other adverbs usually ends at the end of the clause or sentence. Conjunctions generally have a scope over the syntactic unit whose members they coordinate. Some linguistic phenomena (e.g. passive voice or raising) can change scope boundaries in the sentence, thus they were given special attention during the annotation phase.

3.2 Wikipedia Weasels

The chief editors of Wikipedia have drawn the attention of the public to uncertainty issues they call weasel[1]. A word is considered to be a weasel word if it creates an impression that something important has been said, but what is really communicated is vague, misleading, evasive or ambiguous. Weasel words do not give a neutral account of facts, rather, they offer an opinion without any backup or source. The following sentence does not specify the source of information, it is just the vague term *some people* that refers to the holder of this opinion:

> **Some people** claim that this results in a better taste than that of other diet colas (most of which are sweetened with aspartame alone).

Statements with weasel words usually evoke questions such as *Who says that?, Whose opinion is this?* and *How many people think so?*.

Typical instances of weasels can be grouped in the following way (we offer some examples as well):

- Adjectives and adverbs
  - elements referring to uncertainty: *probable, likely, possible, unsure, often, possibly, allegedly, apparently, perhaps*, etc.
  - elements denoting generalization: *widely, traditionally, generally, broadly-accepted, widespread*, etc.
  - qualifiers and superlatives: *global, superior, excellent, immensely, legendary, best, (one of the) largest, most prominent*, etc.
  - elements expressing obviousness: *clearly, obviously, arguably*, etc.
- Auxiliaries
  - *may, might, would, should*, etc.
- Verbs
  - verbs with speculative content and their passive forms: *suggest, question, presume, suspect, indicate, suppose, seem, appear, favor*, etc.
  - passive forms with dummy subjects: *It is claimed that ... It has been mentioned ... It is known ...*
  - *there is / there are* constructions: *There is evidence/concern/indication that...*
- Numerically vague expressions / quantifiers
  - *certain, numerous, many, most, some, much, everyone, few, various, one group of*, etc. *Experts say ... Some people think ... More than 60% percent ...*

[1] http://en.wikipedia.org/wiki/Weasel_word

- Nouns
  - *speculation, proposal, consideration*, etc. *Rumour has it that ... Common sense insists that ...*

However, the use of the above words or grammatical devices does not necessarily entail their being a weasel cue since their use may be justifiable in their contexts.

As the main application goal of weasel detection is to highlight articles which should be improved (by reformulating or adding factual issues), we decided to annotate only weasel cues in Wikipedia articles, but we did not mark their scopes.

During the manual annotation process, the following cue marking principles were employed. Complex verb phrases were annotated as weasel cues since in some cases, both the passive construction and the verb itself are responsible for the weasel. In passive forms with dummy subjects and *there is / there are* constructions, the weasel cue included the grammatical subject (i.e. *it* and *there*) as well. As for numerically vague expressions, the noun phrase containing a quantifier was marked as a weasel cue. If there was no quantifier (in the case of a bare plural), the noun was annotated as a weasel cue. Comparatives and superlatives were annotated together with their article. Anaphoric pronouns referring to a weasel word were also annotated as weasel cues.

4 Task Definitions

Two uncertainty detection tasks (sentence classification and in-sentence hedge scope detection) in two domains (biological publications and Wikipedia articles) with three types of submissions (closed, cross and open) were given to the participants of the CoNLL 2010 Shared Task.

4.1 Detection of Uncertain Sentences

The aim of Task1 was to develop automatic procedures for identifying sentences in texts which contain unreliable or uncertain information. In particular, this task is a binary classification problem, i.e. factual and uncertain sentences have to be distinguished.

As training and evaluation data

- **Task1B**: biological abstracts and full articles (evaluation data contained only full articles) from the BioScope corpus and
- **Task1W**: paragraphs from Wikipedia possibly containing weasel information

were provided. The annotation of weasel/hedge cues was carried out on the phrase level, and sentences containing at least one cue were considered as uncertain, while sentences with no cues were considered as factual. The participating systems had to submit a binary classification (certain vs. uncertain) of the test sentences while marking cues in the submissions was voluntary (but participants were encouraged to do this).

4.2 In-sentence Hedge Scope Resolution

For Task2, in-sentence scope resolvers had to be developed. The training and evaluation data consisted of biological scientific texts, in which instances of speculative spans – that is, keywords and their linguistic scope – were annotated manually. Submissions to Task2 were expected to automatically annotate the cue phrases and the left and right boundaries of their scopes (exactly one scope must be assigned to a cue phrase).

4.3 Evaluation Metrics

The evaluation for Task1 was carried out at the sentence level, i.e. the cue annotations in the sentence were not taken into account. The $F_{\beta=1}$ measure (the harmonic mean of precision and recall) of the uncertain class was employed as the chief evaluation metric.

The Task2 systems were expected to mark cue- and corresponding scope begin/end tags linked together by using some unique IDs. A scope-level $F_{\beta=1}$ measure was used as the chief evaluation metric where true positives were scopes which exactly matched the gold standard cue phrases and gold standard scope boundaries assigned to the cue word. That is, correct scope boundaries with incorrect cue annotation and correct cue words with bad scope boundaries were both treated as errors.

This scope-level metric is very strict. For instance, the requirement of the precise match of the cue phrase is questionable as – from an application point of view – the goal is to find uncertain text spans and the evidence for this is not so important. However, the annotation of cues in datasets is essential for training scope detectors since locating the cues usually precedes the identification of their scope. Hence we decided to incorporate cue matches into the evaluation metric.

Another questionable issue is the strict boundary matching requirement. For example, including or excluding punctuations, citations or some bracketed expressions, like *(see Figure 1)* from a scope is not crucial for an otherwise accurate scope detector. On the other hand, the list of such ignorable phenomena is arguable, especially across domains. Thus, we considered the strict boundary matching to be a straightforward and unambiguous evaluation criterion. Minor issues like those mentioned above could be handled by simple post-processing rules. In conclusion we think that the uncertainty detection community may find more flexible evaluation criteria in the future but the strict scope-level metric is definitely a good starting point for evaluation.

4.4 Closed and Open Challenges

Participants were invited to submit results in different configurations, where systems were allowed to exploit different kinds of annotated resources. The three possible submission categories were:

- **Closed**, where only the labeled and unlabeled data provided for the shared task were allowed, separately for each domain (i.e. biomedical train data for biomedical test set and Wikipedia train data for Wikipedia test set). No further manually crafted resources of uncertainty information (i.e. lists, annotated data, etc.) could be used in any domain. On the other hand, tools exploiting the manual annotation of linguistic phenomena not related to uncertainty (such as POS taggers and parsers trained on labeled corpora) were allowed.

- **Cross**-domain was the same as the closed one but all data provided for the shared task were allowed for both domains (i.e. Wikipedia train data for the biomedical test set, the biomedical train data for Wikipedia test set or a union of Wikipedia and biomedical train data for both test sets).

- **Open**, where any data and/or any additional manually created information and resource (which may be related to uncertainty) were allowed for both domains.

The motivation behind the cross-domain and the open challenges was that in this way, we could assess whether adding extra (i.e. not domain-specific) information to the systems can contribute to the overall performance.

5 Datasets

Training and evaluation corpora were annotated manually for hedge/weasel cues and their scope by two independent linguist annotators. Any differences between the two annotations were later resolved by the chief annotator, who was also responsible for creating the annotation guidelines and training the two annotators. The datasets are freely available[2] for further benchmark experiments at `http://www.inf.u-szeged.hu/rgai/conll2010st`.

Since uncertainty cues play an important role in detecting sentences containing uncertainty, they are tagged in the Task1 datasets as well to enhance training and evaluation of systems.

5.1 Biological Publications

The biological training dataset consisted of the biological part of the BioScope corpus (Vincze et al., 2008), hence it included abstracts from the GENIA corpus, 5 full articles from the functional genomics literature (related to the fruit fly) and 4 articles from the open access BMC Bioinformatics website. The automatic segmentation of the documents was corrected manually and the sentences (14541 in number) were annotated manually for hedge cues and their scopes.

The evaluation dataset was based on 15 biomedical articles downloaded from the publicly available PubMedCentral database, including 5 random articles taken from the *BMC Bioinformatics* journal in October 2009, 5 random articles to which the *drosophila* MeSH term was assigned and 5 random articles having the MeSH terms *human, blood cells* and *transcription factor* (the same terms which were used to create the Genia corpus). These latter ten articles were also published in 2009. The aim of this article selection procedure was to have a theme that was close to the training corpus. The evaluation set contained 5003 sentences, out of which 790 were uncertain. These texts were manually annotated for hedge cues and their scope. To annotate the training and the evaluation datasets, the same annotation principles were applied.

[2] under the Creative Commons Attribute Share Alike license

For both Task1 and Task2, the same dataset was provided, the difference being that for Task1, only hedge cues and sentence-level uncertainty were given, however, for Task2, hedge cues and their scope were marked in the text.

5.2 Wikipedia Datasets

2186 paragraphs collected from Wikipedia archives were also offered as Task1 training data (11111 sentences containing 2484 uncertain ones). The evaluation dataset contained 2346 Wikipedia paragraphs with 9634 sentences, out of which 2234 were uncertain.

For the selection of the Wikipedia paragraphs used to construct the training and evaluation datasets, we exploited the weasel tags added by the editors of the encyclopedia (marking unsupported opinions or expressions of a non-neutral point of view). Each paragraph containing weasel tags (5874 different ones) was extracted from the history dump of English Wikipedia. First, 438 randomly selected paragraphs were manually annotated from this pool then the most frequent cue phrases were collected. Later on, two other sets of Wikipedia paragraphs were gathered on the basis of whether they contained such cue phrases or not. The aim of this sampling procedure was to provide large enough training and evaluation samples containing weasel words and also occurrences of typical weasel words in non-weasel contexts.

Each sentence was annotated manually for weasel cues. Sentences were treated as uncertain if they contained at least one weasel cue, i.e. the scope of weasel words was the entire sentence (which is supposed to be rewritten by Wikipedia editors).

5.3 Unlabeled Data

Unannotated but pre-processed full biological articles (150 articles from the publicly available PubMedCentral database) and 1 million paragraphs from Wikipedia were offered to the participants as well. These datasets did not contain any manual annotation for uncertainty, but their usage permitted data sampling from a large pool of in-domain texts without time-wasting pre-processing tasks (cleaning and sentence splitting).

5.4 Data Format

Both training and evaluation data were released in a custom XML format. For each task, a separate XML file was made available containing the whole document set for the given task. Evaluation datasets were available in the same format as training data without any sentence-level certainty, cue or scope annotations.

The XML format enabled us to provide more detailed information about the documents such as segment boundaries and types (e.g. section titles, figure captions) and it is the straightforward format to represent nested scopes. Nested scopes have overlapping text spans which may contain cues for multiple scopes (there were 1058 occurrences in the training and evaluation datasets together). The XML format utilizes id-references to determine the scope of a given cue. Nested constructions are rather complicated to represent in the standard IOB format, moreover, we did not want to enforce a uniform tokenization.

To support the processing of the data files, reader and writer software modules were developed and offered to the participants for the uCompare (Kano et al., 2009) framework. uCompare provides a universal interface (UIMA) and several text mining and natural language processing tools (tokenizers, POS taggers, syntactic parsers, etc.) for general and biological domains. In this way participants could configure and execute a flexible chain of analyzing tools even with a graphical UI.

6 Submissions and Results

Participants uploaded their results through the shared task website, and the official evaluation was performed centrally. After the evaluation period, the results were published for the participants on the Web. A total of 23 teams participated in the shared task. 22, 16 and 13 teams submitted output for Task1B, Task1W and Task2, respectively.

6.1 Results

Tables 1, 2 and 3 contain the results of the submitted systems for Task1 and Task2. The last name of the first author of the system description paper (published in these proceedings) is used here as a system name[3]. The last column contains the type of submission. The system of Kilicoglu and Bergler (2010) is the only open submission. They adapted their system introduced in Kilicoglu and Bergler (2008) to the datasets of the shared task.

Regarding cross submissions, Zhao et al. (2010) and Ji et al. (2010) managed to achieve a noticeable improvement by exploiting cross-domain

[3] Özgür did not publish a description of her system.

Name	P / R / F	type
Georgescul	72.0 / 51.7 / 60.2	C
Ji	62.7 / 55.3 / 58.7	X
Chen	68.0 / 49.7 / 57.4	C
Morante	80.6 / 44.5 / 57.3	C
Zhang	76.6 / 44.4 / 56.2	C
Zheng	76.3 / 43.6 / 55.5	C
Täckström	78.3 / 42.8 / 55.4	C
Sanchez	68.3 / 46.2 / 55.1	C
Tang	82.3 / 41.4 / 55.0	C
Kilicoglu	67.9 / 46.0 / 54.9	O
Tjong Kim Sang	74.0 / 43.0 / 54.4	C
Clausen	75.1 / 42.0 / 53.9	C
Özgür	59.4 / 47.9 / 53.1	C
Zhou	85.3 / 36.5 / 51.1	C
Li	88.4 / 31.9 / 46.9	C
Prabhakaran	88.0 / 28.4 / 43.0	C
Ji	94.2 / 6.6 / 12.3	C

Table 1: Task1 Wikipedia results (type $\in$ {Closed(C), Cross(X), Open(O)}).

data. Zhao et al. (2010) extended the biological cue word dictionary of their system – using it as a feature for classification – by the frequent cues of the Wikipedia dataset, while Ji et al. (2010) used the union of the two datasets for training (they have reported an improvement from 47.0 to 58.7 on the Wikipedia evaluation set after a post-challenge bugfix).

Name	P / R / F	type
Morante	59.6 / 55.2 / 57.3	C
Rei	56.7 / 54.6 / 55.6	C
Velldal	56.7 / 54.0 / 55.3	C
Kilicoglu	62.5 / 49.5 / 55.2	O
Li	57.4 / 47.9 / 52.2	C
Zhou	45.6 / 43.9 / 44.7	O
Zhou	45.3 / 43.6 / 44.4	C
Zhang	46.0 / 42.9 / 44.4	C
Fernandes	46.0 / 38.0 / 41.6	C
Vlachos	41.2 / 35.9 / 38.4	C
Zhao	34.8 / 41.0 / 37.7	C
Tang	34.5 / 31.8 / 33.1	C
Ji	21.9 / 17.2 / 19.3	C
Täckström	2.3 / 2.0 / 2.1	C

Table 2: Task2 results (type $\in$ {Closed(C), Open(O)}).

Each Task2 and Task1W system achieved a higher precision than recall. There may be two reasons for this. The systems may have applied only reliable patterns, or patterns occurring in the evaluation set may be imperfectly covered by the training datasets. The most intense participation was on Task1B. Here, participants applied various precision/recall trade-off strategies. For instance, Tang et al. (2010) achieved a balanced precision/recall configuration, while Li et al. (2010) achieved third place thanks to their superior precision.

Name	P / R / F	type
Tang	85.0 / 87.7 / 86.4	C
Zhou	86.5 / 85.1 / 85.8	C
Li	90.4 / 81.0 / 85.4	C
Velldal	85.5 / 84.9 / 85.2	C
Vlachos	85.5 / 84.9 / 85.2	C
Täckström	87.1 / 83.4 / 85.2	C
Shimizu	88.1 / 82.3 / 85.1	C
Zhao	83.4 / 84.8 / 84.1	X
Özgür	77.8 / 91.3 / 84.0	C
Rei	83.8 / 84.2 / 84.0	C
Zhang	82.6 / 84.7 / 83.6	C
Kilicoglu	92.1 / 74.9 / 82.6	O
Morante	80.5 / 83.3 / 81.9	X
Morante	81.1 / 82.3 / 81.7	C
Zheng	73.3 / 90.8 / 81.1	C
Tjong Kim Sang	74.3 / 87.1 / 80.2	C
Clausen	79.3 / 80.6 / 80.0	C
Szidarovszky	70.3 / 91.0 / 79.3	C
Georgescul	69.1 / 91.0 / 78.5	C
Zhao	71.0 / 86.6 / 78.0	C
Ji	79.4 / 76.3 / 77.9	C
Chen	74.9 / 79.1 / 76.9	C
Fernandes	70.1 / 71.1 / 70.6	C
Prabhakaran	67.5 / 19.5 / 30.3	X

Table 3: Task1 biological results (type $\in$ {Closed(C), Cross(X), Open(O)}).

Tables 4 and 5 show the cue-level performances, i.e. the F-measure of cue phrase matching where true positives were strict matches. Note that it was optional to submit cue annotations for Task1 (if participants submitted systems for both Task2 and Task1B with cue tagging, only the better score of the two was considered).

It is interesting to see that Morante et al. (2010) who obtained the best results on Task2 achieved a medium-ranked F-measure on the cue-level (e.g. their result on the cue-level is lower by 4% com-

pared to Zhou et al. (2010), while on the scope-level the difference is 13% in the reverse direction), which indicates that the real strength of the system of Morante et al. (2010) is the accurate detection of scope boundaries.

Name	P / R / F
Tang	63.0 / 25.7 / 36.5
Li	76.1 / 21.6 / 33.7
Özgür	28.9 / 14.7 / 19.5
Morante	24.6 / 7.3 / 11.3

Table 4: Wikipedia cue-level results.

Name	P / R / F	type
Tang	81.7 / 81.0 / 81.3	C
Zhou	83.1 / 78.8 / 80.9	C
Li	87.4 / 73.4 / 79.8	C
Rei	81.4 / 77.4 / 79.3	C
Velldal	81.2 / 76.3 / 78.7	C
Zhang	82.1 / 75.3 / 78.5	C
Ji	78.7 / 76.2 / 77.4	C
Morante	78.8 / 74.7 / 76.7	C
Kilicoglu	86.5 / 67.7 / 76.0	O
Vlachos	82.0 / 70.6 / 75.9	C
Zhao	76.7 / 73.9 / 75.3	X
Fernandes	79.2 / 64.7 / 71.2	C
Zhao	63.7 / 74.1 / 68.5	C
Täckström	66.9 / 58.6 / 62.5	C
Özgür	49.1 / 57.8 / 53.1	C

Table 5: Biological cue-level results (type $\in$ {Closed(C), Cross(X), Open(O)}).

6.2 Approaches

The approaches to Task1 fall into two major categories. There were six systems which handled the task as a classical sentence classification problem and employed essentially a bag-of-words feature representation (they are marked as BoW in Table 6). The remaining teams focused on the cue phrases and sought to classify every token if it was a part of a cue phrase, then a sentence was predicted as uncertain if it contained at least one recognized cue phrase. Five systems followed a pure token classification approach (TC) for cue detection while others used sequential labeling techniques (usually Conditional Random Fields) to identify cue phrases in sentences (SL).

The feature set employed in Task1 systems typically consisted of the wordform, its lemma or stem, POS and chunk codes and about the half of the participants constructed features from the dependency and/or constituent parse tree of the sentences as well (see Table 6 for details).

It is interesting to see that the top ranked systems of Task1B followed a sequence labeling approach, while the best systems on Task1W applied a bag-of-words sentence classification. This may be due to the fact that biological sentences have relatively simple patterns. Thus the context of the cue words (token classification-based approaches used features derived from a window of the token in question, thus, they exploited the relationship among the tokens and their contexts) can be utilized while Wikipedia weasels have a diverse nature. Another observation is that the top systems in both Task1B and Task1W are the ones which did not derive features from syntactic parsing.

Each Task2 system was built upon a Task1 system, i.e. they attempted to recognize the scopes for the predicted cue phrases (however, Zhang et al. (2010) have argued that the objective functions of Task1 and Task2 cue detection problems are different because of sentences containing multiple hedge spans).

Most systems regarded multiple cues in a sentence to be independent from each other and formed different classification instances from them. There were three systems which incorporated information about other hedge cues (e.g. their distance) of the sentence into the feature space and Zhang et al. (2010) constructed a cascade system which utilized directly the predicted scopes (it processes cue phrases from left to right) during predicting other scopes in the same sentence.

The identification of the scope for a certain cue was typically carried out by classifying each token in the sentence. Task2 systems differ in the number of class labels used as target and in the machine learning approaches applied. Most systems – following Morante and Daelemans (2009) – used three class labels (F)IRST, (L)AST and NONE. Two participants used four classes by adding (I)NSIDE, while three systems followed a binary classification approach (SCOPE versus NONSCOPE). The systems typically included a post-processing procedure to force scopes to be continuous and to include the cue phrase in question. The machine learning methods applied can be again categorized into sequence labeling (SL)

NAME	approach	machine learner	feature selection	features employed							
				dict	ortho	lemma/stem	POS	chunk	dep	docpart	other
Clausen	BoW	MaxEnt				+	+				hedge cue distance
Chen	BoW	MaxEnt	statistical	+		+	+			+	sentencelength
Fernandes	SL	ETL				+	+	+			
Georgescul	BoW	SVM+paramtuning		+							
Ji	TC	ModAvgPerceptron					+				
Kilicoglu	TC	manual		+		+			+		external dict
Li	SL	CRF+postproc	greedy fwd			+	+	+			
Morante (wiki)	TC	SVM+postproc	statistical	+		+	+	+	+		
Morante (bio)	SL	KNN	statistical	+		+	+	+	+	+	
Prabhakaran	SL	CRF	greedy fwd	+		+	+		+		LevinClass
Rei	SL	CRF		+		+	+		+		
Sánchez	BoW	SVMTreeKernel		+		+	+		+	+	
Shimizu	SL	Bayes Point Machines	GA		+	+	+	+			NEs, unlabeled data
Szidarovszky	SL	CRF	exhaustive	+	+	+					
Täckström	BoW	SVM	greedy fwd			+	+	+	+	+	sentencelength
Tang	SL	CRF,SVMHMM	statistical	+	+	+	+	+			
Tjong Kim Sang	TC	Naive Bayes									
Velldal	TC	MaxEnt	manual	+		+	+		+		
Vlachos	TC	Bayesian LogReg	manual	+		+	+		+		
Zhang	SL	CRF+feature combination	greedy fwd	+		+	+	+	+		NEs
Zhao	SL	CRF	statistical	+		+	+	+			
Zheng	SL	CRF,MaxEnt	manual			+	+	+		+	Constituent Parsing
Zhou	SL	CRF	statistical	+		+	+	+			WordNet

Table 6: System architectures overview for Task1. Approaches: sequence labeling (SL), token classification (TC), bag-of-words model (BoW); Machine learners: Entropy Guided Transformation Learning (ETL), Averaged Perceptron (AP), k-nearest neighbour (KNN); Feature selection: gathering phrases from the training corpus using statistical thresholds (statistical); Features: orthographical information about the token (ortho), lemma or stem of the token (stem), Part-of-Speech codes (POS), syntactic chunk information (chunk), dependency parsing (dep), position inside the document or section information (docpos)

NAME	approach	scope	ML	postproc	tree	dep	multihedge
Fernandes	TC	FL	ETL				
Ji	TC	I	AP			+	
Kilicoglu	HC		manual	+	+	+	
Li	SL	FL	CRF, SVMHMM	+		+	+
Morante	TC	FL	KNN	+		+	
Rei	SL	FIL	manual+CRF	+		+	
Täckström	TC	FI	SVM			+	
Tang	SL	FL	CRF	+	+		+
Velldal	HC		manual			+	
Vlachos	TC	I	Bayesian MaxEnt	+		+	
Zhang	SL	FIL	CRF			+	+
Zhao	SL	FL	CRF	+			
Zhou	SL	FL	CRF	+	+		

Table 7: System architectures overview for Task2. Approaches: sequence labeling (SL), token classification (TC), hand-crafted rules (HC); Machine learners: Entropy Guided Transformation Learning (ETL), Averaged Perceptron (AP), k-nearest neighbour (KNN); The way of identifying scopes: predicting first/last tokens (FL), first/inside/last tokens (FIL), just inside tokens (I); Multiple Hedges: the system applied a mechanism for handling multiple hedges inside a sentence

and token classification (TC) approaches (see Table 7). The feature sets used here are the same as for Task1, extended by several features describing the relationship between the cue phrase and the token in question mostly by describing the dependency path between them.

7 Conclusions

The CoNLL 2010 Shared Task introduced the novel task of uncertainty detection. The challenge consisted of a sentence identification task on uncertainty (Task1) and an in-sentence hedge scope detection task (Task2). In the latter task the goal of automatic systems was to recognize speculative text spans inside sentences.

The relatively high number of participants indicates that the problem is rather interesting for the Natural Language Processing community. We think that this is due to the practical importance of the task for (principally biomedical) applications and because it addresses several open research questions. Although several approaches were introduced by the participants of the shared task and we believe that the ideas described in this proceedings can serve as an excellent starting point for the development of an uncertainty detector, there is a lot of room for improving such systems. The manually annotated datasets and software tools developed for the shared task may act as benchmarks for these future experiments (they are freely available at `http://www.inf.u-szeged.hu/rgai/conll2010st`).

Acknowledgements

The authors would like to thank Joakim Nivre and Lluís Márquez for their useful suggestions, comments and help during the organisation of the shared task.

This work was supported in part by the National Office for Research and Technology (NKTH, `http://www.nkth.gov.hu/`) of the Hungarian government within the framework of the projects TEXTREND, BELAMI and MASZEKER.

References

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.

Wendy W. Chapman, David Chu, and John N. Dowling. 2007. ConText: An Algorithm for Identifying Contextual Features from Clinical Text. In *Proceedings of the ACL Workshop on BioNLP 2007*, pages 81–88.

Mike Conway, Son Doan, and Nigel Collier. 2009. Using Hedges to Enhance a Disease Outbreak Report

Text Mining System. In *Proceedings of the BioNLP 2009 Workshop*, pages 142–143, Boulder, Colorado, June. Association for Computational Linguistics.

Carol Friedman, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. 1994. A General Natural-language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Viola Ganter and Michael Strube. 2009. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176, Suntec, Singapore, August. Association for Computational Linguistics.

Feng Ji, Xipeng Qiu, and Xuanjing Huang. 2010. Detecting Hedge Cues and their Scopes with Average Perceptron. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task*, pages 139–146, Uppsala, Sweden, July. Association for Computational Linguistics.

Yoshinobu Kano, William A. Baumgartner, Luke McCrohon, Sophia Ananiadou, Kevin B. Cohen, Lawrence Hunter, and Jun'ichi Tsujii. 2009. U-Compare: Share and Compare Text Mining Tools with UIMA. *Bioinformatics*, 25(15):1997–1998, August.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 46–53, Columbus, Ohio, June. Association for Computational Linguistics.

Halil Kilicoglu and Sabine Bergler. 2009. Syntactic Dependency Based Heuristics for Biological Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 119–127, Boulder, Colorado, June. Association for Computational Linguistics.

Halil Kilicoglu and Sabine Bergler. 2010. A High-Precision Approach to Detecting Hedges and Their Scopes. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task*, pages 103–110, Uppsala, Sweden, July. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

George Lakoff. 1972. Linguistics and natural logic. In *The Semantics of Natural Language*, pages 545–665, Dordrecht. Reidel.

Xinxin Li, Jianping Shen, Xiang Gao, and Xuan Wang. 2010. Exploiting Rich Features for Detecting Hedges and Their Scope. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task*, pages 36–41, Uppsala, Sweden, July. Association for Computational Linguistics.

Marc Light, Xin Ying Qiu, and 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*, pages 17–24.

Ben Medlock and Ted Briscoe. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the ACL*, pages 992–999, Prague, Czech Republic, June.

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.

Roser Morante, Vincent Van Asch, and Walter Daelemans. 2010. Memory-based Resolution of In-sentence Scopes of Hedge Cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task*, pages 48–55, Uppsala, Sweden, July. Association for Computational Linguistics.

Arzucan Özgür and Dragomir R. Radev. 2009. Detecting Speculations and their Scopes in Scientific Text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1398–1407, Singapore, August. Association for Computational Linguistics.

György Szarvas. 2008. Hedge Classification in Biomedical Texts with a Weakly Supervised Selection of Keywords. In *Proceedings of ACL-08: HLT*, pages 281–289, Columbus, Ohio, June. Association for Computational Linguistics.

Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan, and Shixi Fan. 2010. A Cascade Method for Detecting Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task*, pages 25–29, Uppsala, Sweden, July. Association for Computational Linguistics.

Sofie Van Landeghem, Yvan Saeys, Bernard De Baets, and Yves Van de Peer. 2009. Analyzing Text in Search of Bio-molecular Events: A High-precision Machine Learning Framework. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for*

Shared Task, pages 128–136, Boulder, Colorado, June. Association for Computational Linguistics.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Shaodian Zhang, Hai Zhao, Guodong Zhou, and Baoliang Lu. 2010. Hedge Detection and Scope Finding by Sequence Labeling with Procedural Feature Selection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task*, pages 70–77, Uppsala, Sweden, July. Association for Computational Linguistics.

Qi Zhao, Chengjie Sun, Bingquan Liu, and Yong Cheng. 2010. Learning to Detect Hedges and their Scope Using CRF. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task*, pages 64–69, Uppsala, Sweden, July. Association for Computational Linguistics.

Huiwei Zhou, Xiaoyan Li, Degen Huang, Zezhong Li, and Yuansheng Yang. 2010. Exploiting Multi-Features to Detect Hedges and Their Scope in Biomedical Texts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task*, pages 56–63, Uppsala, Sweden, July. Association for Computational Linguistics.