Learning Local Content Shift Detectors from Document-level Information

Richárd Farkas
Institute for Natural Language Processing
University of Stuttgart
farkas@ims.uni-stuttgart.de

Abstract

Information-oriented document labeling is a special document multi-labeling task where the target labels refer to a specific information instead of the topic of the whole document. These kind of tasks are usually solved by looking up indicator phrases and analyzing their local context to filter false positive matches. Here, we introduce an approach for machine learning *local content shifters* which detects irrelevant local contexts using just the original document-level training labels. We handle content shifters in general, instead of learning a particular language phenomenon detector (e.g. negation or hedging) and form a single system for document labeling and content shift detection. Our empirical results achieved 24% error reduction – compared to supervised baseline methods – on three document labeling tasks.

1 Introduction

There are special document multi-labeling tasks where the target labels refer to a specific piece of information extractable from the document instead of the overall topic of the document. In these kinds of tasks the target information is usually an attribute or relation related to the target entity (usually a person or an organisation) of the document in question, but the task is to assign class labels at the document (entity) level. For example, the smoking habits of the patients are frequently discussed in the textual parts of clinical notes (Uzuner et al., 2008). In this case the task is to find specific information in the text – i.e. the patient in question is a smoker, past smoker, non-smoker – but at the end an application has to assign labels to the documents(patients). Similarly, the soccer club names where a sportsman played for are document(sportman)-level labels in Wikipedia articles expressed by the Wikipedia categories. The target information in these tasks is usually just mentioned in the document and much of the document is irrelevant for this information request in contrast to standard document classification tasks where the goal is to identify the topics of the whole document. On the other hand, they are not a standard information extraction task as the task is to assign class labels to documents, and the training dataset contains labels just at this level. These special tasks lie somewhere between information extraction and document classification and require special approaches to solve them. We will call them *Information-oriented document labeling* throughout this paper. There are several application areas where information-oriented document labels are naturally present in an enormous amount like clinical records, Wikipedia categories and user-generated tags of news.

Previous evaluation campaigns (Uzuner et al., 2008; Pestian et al., 2007; Uzuner, 2009) demonstrated that information-oriented document labeling can be effectively performed by looking up *indicator phrases* which can be gathered by hand, by corpus statistics or in a hybrid way. However these campaigns also highlighted that the analysis of the *local context* of the indicator phrases is crucial. For instance, in the smoking habit detection task there are a few indicator words (e.g. *smokes*, *cigarette*) and the local context of their occurrences in texts should

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 759–770,
Edinburgh, Scotland, UK, July 27–31, 2011.

be analysed to see whether their semantic was radically changed (e.g. they are negated or in a past tense), for instance:

> *The patient has a 20 pack-year smoking history.*
>
> *The patient denies any smoking history.*
>
> *He has a greater than 100 pack year smoking history and quit 9 to 10 years ago.*

We propose a simple but efficient approach for information-oriented document labeling tasks by addressing the automatic detection of language phenomena for a particular task which alters the sense or information content of the indicator phrase's occurrences. For example, they may be logical modifiers (e.g. negation) or modal modifiers (e.g. auxiliaries like *might* and *can*); they may refer to a subject which differs from the target entity of the task (e.g. clinical notes usually contain information about the family history of the patient); or the semantic content of the shifter may change the role of the target span of a text (e.g. a sportsman can play *for* or *against* a particular team). We call these phenomena *content shifters* and the task of identifying them *content shift detection (CSD)*.

Existing CSD approaches focus on a particular class of language phenomena (especially negation or hedging) and use hand-crafted rules (Chapman et al., 2007) or a supervised learning approach that exploits corpora manually annotated at the token-level for a particular type of content shifter (Morante et al., 2009). Moreover higher level applications (like document labeling and information extraction) use a separate CSD module which is developed independently from the target task. We argue that the nature of content shifters is domain and task dependent, so training corpora (at the token-level) are required for content shifters which are important for a particular task but the construction of such training corpora is expensive. Here, we propose an alternative approach which uses only document-level labels.

The input of our system is a training corpus labeled on the document level (e.g. a clinical dataset consisting clinical notes and meta-data about patients). Our approach extracts indicator phrases and trains a CSD jointly. We focus on local content shifters and we analyse just the sentences of indicator phrase occurrences. Our chief assumption is that CSD can be learnt by exploiting the false positive occurrences of indicator phrases in the training dataset. We show that our method performs significantly better than standard document classifiers (which were designed for a slightly different task).

The chief contributions of our work are that (i) we handle the CSD problem in general, so we detect all content shifters instead of focusing on one particular language phenomenon, (ii) we form a single framework for joint CSD and document labeling, (iii) moreover our approach does not require a dedicated annotated training dataset for content shifters.

2 Related Work

Information-oriented document classification tasks were first highlighted in the clinical domain where medical reports contain useful information about the patient in question, but labels are only available at the document (patient) level. The field of clinical NLP has been studied extensively since the 1990s (Larkey and Croft, 1995), but the most recent results are related to the shared task challenges organized relatively recently (Pestian et al., 2007; Uzuner et al., 2008; Uzuner, 2009). For example the first I2B2 challenge in 2006 (Uzuner et al., 2008) focused on the smoking habits of the patient, the CMC challenge in 2007 (Pestian et al., 2007) dealt with the problem of automatically constructing ICD coding systems and the second I2B2 challenge (Uzuner, 2009) addressed the classification of discharge summaries according to the question "Who's obese and what co-morbidities do they have?". These challenges were dominated by entirely or partly rule-based systems that solved the tasks using indicator phrase lookup and incorporated explicit mechanisms for detecting speculation and negation.

Another domain for information-oriented document classification might be Wikipedia, which contains rich information about entities like persons, places or organisations. Some items of information are available about these entities in the form of categories and infoboxes assigned to articles. Automatic document labeling methods can be trained based on these assignments (Schönhofen, 2006), but these labels do not refer to the main theme of the article but

to a certain type of information.

Existing content shift detection approaches focus on a particular class of language phenomena, especially on negation and hedge recognitions. Available tools work mainly on clinical and biological domains. The first systems were fully hand-crafted (Light et al., 2004; Friedman et al., 1994; Chapman et al., 2007) without any empirical evaluation on a dedicated corpus. Recently, there have been several corpora published with manual sentence-, event- or token-level annotation for negation, certainty and factuality in the biological (Medlock and Briscoe, 2007; Vincze et al., 2008), newswire (Strassel et al., 2008; Sauri and Pustejovsky, 2009) and encyclopedical (Farkas et al., 2010) domains.

Exploiting these corpora, machine learning models were also developed. Solving the sentence-level task, Medlock and Briscoe (2007) used single words as input features in order to classify sentences from biological articles as speculative or non-speculative. Szarvas (2008) extended their methodology to use n-gram features and a semi-supervised selection of the keyword features. Ganter and Strube (2009) proposed an approach for the automatic detection of sentences containing uncertainty based on Wikipedia weasel tags and syntactic patterns. For in-sentence negation and speculation detection, Morante et al. (2009) developed scope – i.e. content shifted text spans – detectors for negation and speculation following a supervised sequence labeling approach, while Özgür and Radev (2009) developed a rule-based system that exploits syntactic patterns. The goal of the CoNLL 2010 Shared Task (Farkas et al., 2010) was to develop linguistic scope detectors as well. The participants usually followed a supervised sequence labeling approach or used a rule-based system that exploits syntactic patterns. The approach of classifying identified events into whether they fall under negation or speculation was followed by Sauri and Pustejovsky (2009) and the participants of the BioNLP'09 Shared Task (Kim et al., 2009). Here the systems investigated the syntax path between the event trigger and a cue word (which came from a small lexicon) (Kilicoglu and Bergler, 2009; Aramaki et al., 2009).

Our approach differs from the previous works fundamentally. We deal with the two tasks (information-oriented document classification and content shift detection) together and introduce a co-learning approach for them. Our approach handles content shifters in a data-driven and generalized way i.e. it is not specialized for a certain class of language phenomena. Instead it tries to recognize task-specific syntactic and semantic patterns which are responsible for semantic changes or irrelevance. In addition, we have no access to a gold-standard sentence-level or in-sentence-level annotation but exploit document-level ones.

3 Tasks and Datasets

Before introducing our approach in detail we describe three tasks and datasets which were used in our experiments in order to give an insight into the challenges of the information-oriented document labeling tasks. Table 1 summarizes the key statistical figures (the number of documents in the corpora, the size of the label sets along with the average number of tokens and label assignments per document) of the datasets used for the experimental evaluations.

Table 1: The datasets used in our experiments.

	CMC	Obes	Soccer
domain	clinical	clinical	encycl.
|train|	978	730	4850
|eval|	976	507	1736
#token/d	25	1387	389
#labels	45	16	12
#label/d	1.24	4.37	1.23

The CMC ICD Coding Dataset was originally prepared for a shared task challenge organized by the Computational Medicine Center (CMC) in Cincinatti, Ohio in 2007 (Pestian et al., 2007). It contains radiology reports along with document-level International Classification of Diseases (ICD) codes given by three human experts. ICD is a coding of diseases, signs, symptoms and abnormal findings. In our experiments we used the train/evaluation split of the shared task. The ICD coding guide states that negative or uncertain diagnosis should not be coded in any case.

The corpus contains very short documents. For instance, the document

HISTORY: Left lower chest pain. Rule-out

pneumonia. IMPRESSION: Normal chest.

has one label *786.50 (cough)* as *486 (pneumonia)* is ruled out.

The main conclusion of the shared task in 2007 was that simple rule-based systems generally outperform bag-of-words-based machine learning models. The rules were extracted from ICD guidelines and/or from the training corpus using simple statistical measures, then they were checked or extended manually. Several systems of the challenge employed a negation and speculation detection submodule. The (manually highly fine-tuned) top systems of the CMC shared task achieved an F-measure of 88-89 (Pestian et al., 2007; Farkas and Szarvas, 2008).

The I2B2 Obesity Dataset was also the subject of a clinical natural language processing shared task. The challenge in 2008 focused on analyzing clinical discharge summary texts and addressed the following question: "Who is obese and what co-morbidities do they have?" (Uzuner, 2009). Target diseases (document labels) included obesity and its 15 most frequent co-morbidities exhibited by patients. In our experiments, we used the same train/evaluation split as that of the shared task. Here a special aspect of the corpus is that the documents are semi-structured, i.e. they contain headings like *discharge medications* and *admit diagnosis*. By pasting the given heading to the beginning of each sentence, we incorporated it into the local context. The top performing systems of the shared task employed mostly hand-crafted rules for indicator selection and for negation and uncertainty detection as well. They achieved an F-measure[1] of 96-97 (Uzuner, 2009; Solt et al., 2009).

Wikipedia Soccer Dataset. We constructed a corpus based on Wikipedia articles and categories[2]. The categories assigned to Wikipedia articles can be regarded as labels (for example, the labels of *David Beckham* in the Wikipedia are *English people*, *expatriate soccer player*, *male model* and *A.C. Milan player*, *Manchester United player*). Based on the categories of Wikipedia, classifiers can be trained to tag unlabeled texts or even add missing category assignments to Wikipedia (Schönhofen, 2006).

For a case study we focused on learning English soccer clubs that a given sportsman played for. Note that this task is an information-oriented document labeling task as the clubs for which a sportsman played are usually just mentioned (especially for smaller clubs) in the article of a player. The Wikipedia category *Footballers in England by club* contains 408 subcategories (for the present and past). We selected the best known clubs (where the category label for the club is assigned to more than 500 player pages). Each article referring to a player having a category assignment to these clubs was downloaded and the textual parts were extracted. Then a random 3:1 train:evaluation split of the document set was used.

4 Document-labeling with CSD

We introduce here an iterative solution which selects indicator phrases and trains a content shift detector at the same time. Our focus will be on *multi-label document classification* tasks where multiple class labels can be assigned to a single document. In this study we will not deal with the modeling of inter-label dependencies, so binary (positive versus negative) and multi-class document classifications (where exactly one label has to be assigned to a single document) can be regarded as special cases of this multi-label classification problem. Our resulting multi-label model is then a set of binary classifiers – "assign a label" classifiers for each class label – and the final prediction on a document is simply the union of the labels forecasted by the individual classifiers.

Our key assumption in the multi-label environment is that while indicator phrases have to be selected on a per class basis, the content shifters can be learnt in a class-independent (aggregated) way i.e. we can assume that within one task, each class label belongs to a given semantic domain (determined by the task), thus the content shifters for their indicator phrases are the same. This approach provides an adequate amount of training samples for content shift detector learning.

[1] Using the definitions of the challenge, the evaluation metric applied here is the micro F-measure of the textual task on the YES versus every other class.

[2] The dataset is available as the supplementary material.

Table 2: Example feature representation of local contexts of *Arsenal*. The prefix NP stands for the lemma features from the deepest noun phrase; D,DR and DEP marks the lemmas, roles and their combination in the dependency path, respectively; SUBJ and SUBJD denote the lemmas and dependency roles on the "subject path", respectively.

His brother, Paul had a long career at Newcastle. (sentenceId=1, indicator=Newcastle)	
bag-of-word features	syntax-based features
he, brother, Paul, have, a, long, career, at	NP#a, NP#long, NP#career, NP#at D#career, D#have, DR#prepat, DR#dobj, DEP#career#prepat, DEP#have#dobj SUBJ#brother, SUBJ#Paul, SUBJ#he, SUBJD#he#poss
He was born in Gosforth, Newcastle and played for Arsenal. (sentenceId=2, indicator=Arsenal)	
bag-of-word features	syntax-based features
he, be, bear, in, Gosforth, Newcastle, and, play, for	D#play, DR#prepfor, DEP#play#prepfor SUBJ#he

4.1 Learning Content Shift Detectors

The key idea behind our approach is that a training corpus for task-specific content shifter learning can be automatically generated by exploiting the occurrences of indicators in various contexts. The local context of an indicator is assumed to have altered if it yields a false positive document-level prediction. More precisely, a training dataset can be constructed for learning a content shift detector in a way that the instances are the local contexts of each occurrence of indicator phrases in the training document set. The instances of this content shifter training dataset are then labeled as `non-altered` when the indicated label is among the gold-standard labels of the document in question or is labeled as `altered` otherwise. On this dataset, arbitrary binary classification models (S) can be trained.

As a feature representation of a local context of an indicator phrase, the bag-of-words of the sentence instance (excluding the indicator phrase itself) was used at the beginning. Our preliminary experiments showed that the tokens of the sentence after the indicator played a negligible role, hence we represented contexts just by tokens before the indicator.

Features concerning the syntactic context of the given indicator were also investigated. For this, we extended the feature set with features derived from the constituent and dependency parses of the sentence[3]. First, the deepest noun phrase which includes the indicator phrase was identified, then all lemmas from its subtree were gathered. From the dependency parse, the lemmas and dependency labels on the directed path from the indicator to the root node (main path) were extracted. The directed paths branching from this main path starting with `subject` dependency were also used for feature extraction (note that these walk in opposite direction to that of the main path). The intuition of the latter was that the subject of the given information – as it can differ from the target entity of extraction – is of great importance. We note that we recognize the in-sentence subject and employing a co-reference module would probably increase the value of these features.

Table 2 exemplifies the feature representation of local contexts of the *Newcastle* and *Arsenal* indicators for the Wikipedia soccer task. In both sentences, a naive system would extract *Newcastle* as false positives. We want to learn content shifters from them along with the true positive match of *Arsenal* in sentence 2. From the first example the CSD could learn even that the bag-of-word contains *brother* or the `SUBJ=`*brother*. However, in the second example, the bag-of-word representation is not sufficient to learn that the local context of *Newcastle* is `altered` because it is the subset of the bag-of-word representation of *Arsenal*'s `non-altered` local context. In this case the syntactic context representation can help and in our CSD `DEP=`*play#prepfor* gets high weight for the `non-altered` class.

[3]We parse only the sentences which contain indicator phrase which makes these features computable in reasonable time even on bigger document sets.

4.2 Co-learning of Indicator Selection and CSD

If document labels are available at training time, an iterative approach can be used to learn the local content shift detector and the indicator phrases as well. The training phase of this procedure (see Algorithm 1) has two outputs, namely the set of indicator phrases for each label I and the content shift detector S which is a binary function for determining whether the sense of an indicator in a particular local context is being altered. Good indicator phrases are those that identify the class label in question when they are present. In each step of the iteration we select indicator phrases $I[l]$ for each label l based on the actual state of the document set D'. Using these $I[l]$s we train a CSD S. Then we apply it to the original dataset D and we delete each local context from the documents which was predicted to be altered by S.

Algorithm 1 Co-learning of labels and CSD

Input: L class labels, D labeled training documents
  $D' \leftarrow D$
  repeat
    for all $l \in L$ **do**
      $I[l] \leftarrow$ indicatorSelection(D', l)
    end for
    $S \leftarrow$ learnCSD(D', I)
    $D' \leftarrow$ removeAlteredParts(D, S)
  until convergence
  return I, S

The indicator selection and content shifter learning phases can form an iterative process. The better the selected indicators are, the better the content shift detectors can be learnt. By applying the content shift detector to each token of the documents, each part of the texts lying within the scope of a content shifter can be removed[4]. By using such a cleaned training document set (D'), better indicators can be selected. These steps can be repeated until some convergence criterion is reached. In our experiments we simply used a fixed iteration number to gain an insight into the behavior of the approach.

[4]In our first experiments introduced here, we removed the parts of the documents classified as altered. Instead of removing these parts they may be marked and then different features may be extracted from them.

Algorithm 2 Document labeling with CSD

Input: d document, I indicator sets, S CSD
  pred $\leftarrow \emptyset$
  for all $l \in L$ **do**
    for all $o \in$ occurrences($d, I[l]$) **do**
      if not altered(o, S) **then**
        pred $\leftarrow$ pred $\cup\, l$
      end if
    end for
  end for
  return pred

The prediction procedure of the approach (see Algorithm 2) then looks for occurrences of the indicator phrases in the text and checks whether they are altered in a certain local context. A non-altered indicator directly assigns a class label without any global consistency check on assigned labels.

We note here, that the local relationship among tokens (i.e. the local context) may be taken into account by incorporating this information directly into the feature space of a document classifier (as an alternative of our co-learning procedure), but the number of features would exponentially increase and submodels for each indicator phrases should be learnt which would made such a classification task intractable.

4.3 Indicator Phrase Selection

Indicator phrases are sequences of tokens whose presence implies the positive class. We aimed to extract phrases with the length of 1,2 or 3 (and we used exact matching after lemmatisation). There are several possible ways of developing indicator selection algorithms. One way is to treat it as a special feature selection procedure where the goal is to select a set of features (uni-, bi-, trigrams of a bag-of-word model) which achieves high recall along with moderate precision as false positives are expected to be eliminated by the local CSD in our two-step approach. Indicator selectors can be even derived from most classifiers which are based on feature weighting (like MaxEnt and AvgPerceptron) or feature ranking (like rule-based classifiers)[5] as well. However indicator selection is not the focus of this

[5]A derivation is more complicated or unfeasible for example-based classifiers like SVMs.

Table 3: Results obtained for local content shift detection in a precision/recall/F-measure format.

		CMC	Obesity	Soccer
Trained	BoW	90.7 / 60.7 / 72.7	82.1 / 35.4 / 49.4	75.0 / 70.6 / 72.7
	BoW+syntactic	88.3 / 60.2 / 71.6	84.4 / 33.3 / 47.8	81.0 / 78.9 / 79.9
Hand-crafted	CSSDB	94.7 / 53.3 / 68.2	42.0 / 57.9 / 48.7	36.8 / 9.8 / 15.5
	in-sentence	80.7 / 65.2 / 72.2	70.5 / 40.5 / 51.5	N/A

work.

For our experiments, a feature evaluation-based greedy algorithm was employed to select the set of indicators from the pool of token uni- and bigrams. The aim of the the indicator selection here is to cover each positive documents while introducing a relatively small amount of false positives. The greedy algorithm iteratively selects the 1-best phrase according to a feature evaluation metric based on the actual state of covered documents and adds it to the indicator phrase set. The process is iterated while the score – in terms of the applied feature evaluation metric – of the 1-best phrase is above a threshold t. The quality of the selected indicator set is highly dependent on the stopping threshold t, but as individual feature evaluation functions are very fast and the number of good indicators is usually low (4-5), the whole greedy indicator selection is fast, hence t can be fine-tuned without overfitting on the training sets employing a cross-validation procedure. As a feature evaluation metric we employed $p(+|f)$ the probability of the positive class "+" conditioned on the presence of a feature f because preliminary experiments did not show any significant advances for more complex metrics.

5 Experiments

Experiments were carried out on the three datasets introduced in Section 3 with local content shift detection as an individual task and also to investigate its added value to information-oriented document labeling.

In our experiments, we applied the sentence splitter and lemmatizer implementation of the MorphAdorner package[6] and the Stanford tokenizer and lexicalized PCFG parser (Klein and Manning, 2003)[7].

5.1 Content Shifter Learning Results

In order to evaluate content shift detection as an individual task, a set of indicator phrases have to be fixed as an input to the CSD. We used manually collected indicator phrases for each label for each dataset. We utilized the terms of Farkas and Szarvas (2008) and Farkas et al. (2009) collected for the CMC and Obesity datasets, respectively and club names for the Soccer dataset in our first branch of experiments. Note that the clinical term sets here have been manually fine-tuned as they were developed for participating systems of the shared tasks of the corpora.

Based on the occurrences of these fixed indicator phrases, CSD **train**ing datasets were built from the local contexts of the three datasets and binary classification was carried out by using MaxEnt. Table 3 shows the results achieved by the learnt CSDs using the bag-of-word feature representation (row 1) along with the ones obtained by the feature set that was extended with syntactic patterns (raw 2). Here, the precision/recall/F-measure values measure how many false positive matches of the indicator phrases can be recognized (the F-measure of the `altered` class), i.e. here, the true positives are local contexts of an indicator phrase which do not indicate a document label in the evaluation set and the local content shift detector predicted it to be `altered`.

For comparison purposes, we employed **manually developed CSDs** which were fine-tuned for the medical shared task datasets. Row 3 of Table 3 (we refer to it as *content shifted sentence detection baseline (CSSDB)* later on) shows the results archived by the method which predicts every sentence to be `altered` which contain any *cue phrases* for negation, modality and different experiencer. Note that off-the-shelf tools are available just for these types of content shifters. We collected cue phrases for such a content shifted sentence detection from the

[6] `morphadorner.northwestern.edu/`

[7] The JAVA implementation of the entire framework and dataset adapters can be found as the supplementary material

works of Chapman et al. (2007), Light et al. (2004) and Vincze et al. (2008) and from the experiments of Farkas and Szarvas (2008) and Farkas et al. (2009).

For the CMC and Obesity tasks, hand-crafted in-sentence CSDs were also available (Farkas and Szarvas, 2008; Farkas et al., 2009), i.e. they apply heuristics – which usually tries to recognise clause boundaries – for determining the scope of a negaton/modality cue. This CSD is more fine-grained than the sentence-level one as here a part of a sentence can be detected as `altered` while other parts as `non-altered`. The results of these detectors – two different CSDs, both highly fine-tuned for the corresponding shared task – are listed in the last row of Table 3.

On the CMC dataset, our machine learning approach identified mostly negation and speculation expressions as **content shifters**; the top weighted features for the positive class of the MaxEnt model were *no, without, may* and *vs*. They can filter out false positive matches like

Hyperinflated without focal pneumonia..

On the Obesity dataset, similar content shifters were learnt along with references to family members (like the terms *mother* and *uncle*, and the *family history* header). The significance of these types of content shifters may be illustrated by the following sentence:

History of hypertension in mother and sister.

The soccer task highlighted totally different content shifters which is also the reason for the poor performance of CSSDB. The mention of a club name which the person in question did not play for (false positives) is usually a rival club, club of an unsuccessful negotiation or club which was managed by the footballer after his retirement. For example:

His last game was against Chelsea at Stamford Bridge.

He was a coach at United during his son's playing career.

Summing up, the machine learnt CSDs proved to be competitive with the manually fine-tuned CSD on the three datasets. Table 3 shows that learnt CSDs were able to eliminate a significant amount of false positive indicator phrase matches on each of the three datasets. The hand-crafted CSDs developed for the medical texts certainly work poorly (an F-score of 15.5) on the Soccer dataset as content shifters different from negation, hedge and experiencer are useful there. On the other hand, the content shifters could be learnt on this dataset by our CSD approach (achieving F-score of 79.9). In the clinical corpora, the features from the syntactic parses just confused the system, but they proved to be useful on the Soccer corpus. Here, the dependency parse achieved improvements in terms of both precision and recall (the number of true positives increased by 137) which can be mainly attributed to the prepositions *against* and *over*. The reason why it did not advance on the clinical corpora is probably the domain difference between the training corpus of the parsers and the target texts, i.e. the parsers trained on the Wall Street Journal could not build adequate dependency parses on clinical notes.

As a final comparison we investigated the manually annotated BioScope corpus (Vincze et al., 2008) as a CSD. The CMC corpus is included in the BioScope corpus where text spans in the in-sentence scope of speculation and negation were annotated. We used this manual annotation as an oracle CSD and got an F-measure of 75.2 (which is significantly higher than the scores 72.2 and 72.7 archived by the hand-crafted and trained CSD respectively). This score can be regarded as an upper bound for the amount of false positive indicator matches that can be fixed by local speculation and negation detectors. The remaining false positives are not covered by the linguistically motivated annotations of BioScope, i.e. false positives recognizable by domain knowledge (e.g. coding symptoms should be omitted when a certain diagnosis that is connected with the symptom in question is present in the document) are not marked.

Our **error analysis** revealed that most of the errors of the learnt CSDs is due to the lack of semantic link between lexical units. For instance, on the Soccer dataset it could learn that the token *coach* occuring in the sentence in question indicates an `altered` content, but it was not able to recognise this for *trainer*. The reason for that is simple, the ratio of occurrences of *trainer*:*coach* is 5:95 in the

Table 4: Results obtained by document multi-labeling algorithms in a precision/recall/F-measure format.

rowID				CMC	Obesity	Soccer
1	Baseline	SVM	with CSSDB	87.7 / 76.7 / 81.8	90.0 / 81.3 / 85.4	92.2 / 75.1 / 82.8
2		MaxEnt		92.2 / 72.2 / 81.0	91.4 / 87.6 / 89.4	92.2 / 77.4 / 84.2
3		PART		83.9 / 80.6 / 82.2	87.3 / 86.4 / 86.8	81.2 / 77.0 / 79.0
4	Indicator Selection	without CSD		78.0 / 85.1 / 81.4	89.2 / 93.6 / 91.3	84.4 / 83.7 / 84.1
5		with CSSDB		79.0 / 84.1 / 81.4	94.8 / 86.6 / 91.1	85.2 / 85.5 / 85.3
6		with learnt CSD		83.1 / 83.2 / 83.2	91.7 / 92.9 / 92.3	91.7 / 85.2 / **88.3**
7		after 3 iterations		82.4 / 86.8 / **84.6**	92.6 / 95.4 / **94.0**	92.5 / 84.0 / 88.0
8		after 10 iterations		82.4 / 86.8 / 84.6	92.7 / 95.4 / 94.0	92.5 / 84.0 / 88.0
9	Baseline	MaxEnt with learnt CSD		89.9 / 77.0 / 83.0	91.9 / 90.4 / 91.1	95.0 / 78.7 / 86.1

training corpus. Increasing the training size may be a simple way to overcome this shortcoming. Note that increasing the number of labels (e.g. introducing more soccer clubs in the Soccer task) would also directly increase the size of training dataset as we use the occurrences of the indicator phrases belonging to each of the labels for training a CSD. The solution for the rare cases would require the explicit handling of semantic relatedness (by utilising existing semantic resources or trying to automatically identify task-specific relations).

5.2 Document Labeling Results

The second branch of experiments investigated the added value of CSDs in information-oriented document labeling tasks. Table 4 summarizes the results we got on the three datasets using the micro-averaged $F_{\beta=1}$ of assigned labels (positive class).

As **baseline** systems we trained binary SVMs with a linear kernel, MaxEnts and PARTs – a rule-learner classification algorithm (Frank and Witten, 1998) – for each label using the bag-of-word representation of the documents (implementations of SVMligth (Joachims, 1999), MALLET (McCallum, 2002) and WEKA (Witten and Frank, 1999) were used). The first two learners are popular choices for document classification, while the third is similar to our simple indicator selection procedure. We did not tuned the parameters of the classifiers, we used the default ones everywhere.

To have a fair comparison, we applied to preprocessing steps on dataset of these document classifiers. First, we removed from the training and evaluation raw documents which were predicted to be `altered` by CSSDB. Second, as our indicator selection phrase can be regarded as a special feature selection method, we carried out an Information Gain-based feature selection (keeping the 500 best-rated features proved to be the best solution) on the bag-of-word representation of the documents. The effect of these two preprocessing steps varied among datasets. It improved the F-score of the MaxEnt baseline document classifier by 20%, 2% and 3% on the Obesity, CMC and Soccer datasets, respectively (the F-measures of Table 4 are the values we got by employing pre-processing).

The **indicator selection** results presented in the rows 4-8 of Table 4 made use of the $p(+|f)$-based indicator selector with a five-fold-cross-validated stopping threshold t (introduced in Section 4.3). Row 4 contains the results of using the selected indicators without any CSD. Indicator selection with the CSSDB was applied for the 5th row. Rows 6-8 of Table 4 show the results obtained after one, three and ten iterations of the full learning algorithm (see Algorithm 1). For training the CSD, we employed MaxEnt as a binary classifier for detecting `altered` local contexts and we used the basic BoW feature representation for the clinical tasks while the extended (BoW+syntactic) one for the Soccer dataset.

In the final experiment (the last row of Table 4)) we investigated whether the learnt content shift detector can be applied as a general "**document cleaner**" tool. For this, we trained the baseline MaxEnt document classifier with feature selection on documents from which the text spans predicted to be `altered` by the learnt CSD in the tenth iteration were removed. This means that the systems

used in row 2 and row 9 differ only in the applied document cleaner pre-processing steps (the first one applied the CSSDB while the latter one employed the the learnt CSD).

The difference between the best baseline and the indicator selector with learnt CSD and between the best baseline and the document classifier with learnt CSD were statistically significant[8] on each dataset. The difference between the predictions after the 1st and 3rd iterations were statistically significant on the CMC and the Obesity corpora but not significant on the Soccer dataset. The difference between the 3th and 10th iterations were not significant in either case. Our co-learning method which integrated the document-labeling and CSD tasks significantly outperformed the baseline approaches – which use separate document cleaning and document labeling steps – on the three datasets.

On the clinical domains the automatically selected indicators were disease names, symptom names (e.g. *high blood pressure*), their spelling variants, synonyms (like *hypertension*) and their abbreviations (e.g *htn*). On the soccer domain club names, synonyms (like *The Saints*) and stadium names (e.g. *Old Trafford*) were selected. A label was indicated by 3-4 indicator phrases.

Note that in these information-oriented document multi-labeling tasks simple indicator selection-based document labelers alone achieved results comparable to the bag-of-words-based classifiers. The learnt content shift detectors led to an average improvement of 3.6% in the F-measure (i.e. a 24% error reduction). The effect of further iterations is various. As Table 4 shows, three iterations brought an increase on the CMC and Obesity datasets but not on the Soccer corpus. After a few iterations the set of indicator phrases and the content shift detector did not change substantially. The results achieved by the MaxEnt document classifier employing the ”cleaned” training documents (last row of Table 4) are significantly better (an average improvement of 1.9% in the F-measure and 12% error reduction) than those by the CSSDB (row 2) but the indicator selector approach performed even better.

[8] According to McNemar's test with P-value of 0.001

6 Conclusions

In this paper, we dealt with information-oriented document labeling tasks and investigated machine learning approaches for local content shift detectors from document-level labels. We demonstrated experimentally that a significant amount of false positive matches of indicator phrases can be recognized by trained content shift detectors. Our trained CSD does not use any task or domain specific knowledge and exploits the false and true positive matches of indicator phrases, i.e. it uses only document-level annotation. This task-independent approach achieved competitive results with CSDs which were manually fine-tuned for particular datasets. The empirical results also support the idea of generalized local CSD (false positive removal) opposite to developing independent CSD for particular language phenomena (like negation and speculation).

A co-learning framework for training local content shift detectors and indicator selection was introduced as well. Our method integrates document classification and CSD learning, which are traditionally used as independent submodules of applications. Experiments on three information-oriented document-labeling datasets – from two application areas – with simple indicator selection and syntactic parse-based content shifter learning were performed and the results show a clear improvement over the bag-of-word-based document classification baseline approaches.

However, the proposed content shift detector learning approach is tailored for information-oriented document labeling tasks, i.e. it performs well when not too many and reliable indicator phrases are present. In the future, we plan to investigate and extend the framework for the general document classification task where many indicators with complex relationships among them determine the labels of a document but local content shifters can play an important role.

Acknowledgements

This work was partially founded by the Research Group on Artificial Intelligence of the Hungarian Academy of Sciences. Richárd Farkas was also funded by Deutsche Forschungsgemeinschaft grant SFB 732.

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.

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.

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based icd-9-cm coding systems. *BMC Bioinformatics*, 9(Suppl 3):S10.

Richárd Farkas, György Szarvas, István Hegedüs, Attila Almási, Veronika Vincze, Róbert Ormándi, and Róbert Busa-Fekete. 2009. Semi-automated construction of decision rules to predict morbidities from clinical texts. *Journal of the American Medical Informatics Association*, 16:601–605.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 1–12.

Eibe Frank and Ian H. Witten. 1998. Generating accurate rule sets without global optimization. In *Proc. of Fifteenth International Conference on Machine Learning*, pages 144–151.

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.

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

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

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.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st ACL*, pages 423–430.

Leah S. Larkey and W. Bruce Croft. 1995. Automatic assignment of icd9 codes to discharge summaries. Technical report.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proc. of Biolink 2004, Linking Biological Literature, Ontologies and Databases (HLT-NAACL Workshop:)*, pages 17–24.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the ACL*, pages 992–999, June.

Roser Morante, V. Van Asch, and A. van den Bosch. 2009. Joint memory-based learning of syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*, pages 25–30.

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.

John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the ACL Workshop on BioNLP 2007*, pages 97–104.

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

Peter Schönhofen. 2006. Identifying document topics using the wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 456–462.

Illés Solt, Domonkos Tikk, Viktor Gál, and Zsolt Tivadar Kardkovács. 2009. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. *J. Am. Med. Inform. Assoc.*, 16:580–584, jul.

Stephanie Strassel, Mark A. Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *LREC*.

György Szarvas. 2008. Hedge Classification in Biomedical Texts with a Weakly Supervised Selection of Keywords. In *Proceedings of ACL-08*, pages 281–289.

O. Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying Patient Smoking Status from Medical Discharge Records. *Journal of American Medical Informatics Association*, 15(1):14–24.

Ozlem Uzuner. 2009. Recognizing obesity and comorbidities in sparse data. *Journal of American Medical Informatics Association*, 16(4):561–70.

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.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.