Towards Gene Recognition from Rare and Ambiguous Abbreviations using a Filtering Approach

Matthias Hartung*, Roman Klinger*, Matthias Zwick‡ and Philipp Cimiano*

*Semantic Computing Group
Cognitive Interaction Technology –
Center of Excellence (CIT-EC)
Bielefeld University
33615 Bielefeld, Germany
{mhartung,rklinger,cimiano}
@cit-ec.uni-bielefeld.de

‡Research Networking
Boehringer Ingelheim Pharma GmbH
Birkendorfer Str. 65
88397 Biberach, Germany
matthias.zwick
@boehringer-ingelheim.com

Abstract

Retrieving information about highly ambiguous gene/protein homonyms is a challenge, in particular where their non-protein meanings are more frequent than their protein meaning (*e. g.*, *SAH* or *HF*). Due to their limited coverage in common benchmarking data sets, the performance of existing gene/protein recognition tools on these problematic cases is hard to assess.

We uniformly sample a corpus of eight ambiguous gene/protein abbreviations from MEDLINE® and provide manual annotations for each mention of these abbreviations.[1] Based on this resource, we show that available gene recognition tools such as conditional random fields (CRF) trained on BioCreative 2 NER data or GNAT tend to underperform on this phenomenon.

We propose to extend existing gene recognition approaches by combining a CRF and a support vector machine. In a cross-entity evaluation and without taking any entity-specific information into account, our model achieves a gain of 6 points F_1-Measure over our best baseline which checks for the occurrence of a long form of the abbreviation and more than 9 points over all existing tools investigated.

1 Introduction

In pharmaceutical research, a common task is to gather all relevant information about a gene, *e. g.*, from published articles or abstracts. The task of recognizing the mentions of genes or proteins can be understood as the classification problem to decide whether the entity of interest denotes a gene/protein or something else. For highly ambiguous short names, this task can be particularly challenging. Consider, for instance, the gene *acyl-CoA synthetase medium-chain family member 3* which has synonyms *protein SA homolog* or *SA hypertension-associated homolog*, among others, with abbreviations *ACSM3*, and *SAH*.[2] Standard thesaurus-based search engines would retrieve results where *SAH* denotes the gene/protein of interest, but also occurrences in which it denotes other proteins (*e. g.*, *ATX1 antioxidant protein 1 homolog*[3]) or entities from semantic classes other than genes/proteins (*e. g.*, the symptom *sub-arachnoid hemorrhage*).

For an abbreviation such as *SAH*, the use as denoting a symptom or another semantic class different from genes/proteins is more frequent by a factor of 70 compared to protein-denoting mentions according to our corpus analysis, such that the retrieval precision for *acyl-CoA synthetase* by the occurrence of the synonym *SAH* is only about 0.01, which is totally unacceptable for practical applications.

In this paper, we discuss the specific challenge of recognizing such highly ambiguous abbreviations. We consider eight entities and show that common corpora for gene/protein recognition are of limited value for their investigation. The abbreviations we consider are SAH, MOX, PLS, CLU, CLI, HF, AHR and COPD (*cf.* Table 1). Based on a sample from MEDLINE[4], we show that these names do actually occur in biomedical text, but are underrepresented in corpora typically used for benchmarking and developing gene/protein recognition approaches.

[1] The annotated corpus is available for future research at http://dx.doi.org/10.4119/unibi/2673424.

[2] http://www.ncbi.nlm.nih.gov/gene/6296

[3] http://www.ncbi.nlm.nih.gov/gene/443451

[4] http://www.nlm.nih.gov/pubs/factsheets/medline.html

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 118–127,
Baltimore, Maryland USA, June 26-27 2014. ©2014 Association for Computational Linguistics

Synonym	Other names	Other meaning	EntrezGene ID
SAH	acyl-CoA synthetase medium-chain family member 3; ACSM3	subarachnoid hemorrhage; S-Adenosyl-L-homocysteine hydrolase	6296
MOX	monooxygenase, DBH-like 1	moxifloxacin; methylparaoxon	26002
PLS	POLARIS	partial least squares; primary lateral sclerosis	3770598
CLU	clusterin; CLI	covalent linkage unit	1191
CLI	clusterin; CLU	clindamycin	1191
HF	complement factor H; CFH	high frequency; heart failure; Hartree-Fock	3075
AHR	aryl hydrocarbon receptor; bHLHe76	airway hyperreactivity	196
COPD	archain 1; ARCN1; coatomer protein complex, subunit delta	Chronic Obstructive Pulmonary Disease	22819; 372

Table 1: The eight synonyms for genes/proteins which are subject of analysis in this paper and their long names together with frequent other meanings.

We propose a machine learning-based filtering approach to detect whether a mention in question actually denotes a gene/protein or not and show that for the eight highly ambiguous abbreviations that we consider, the performance of our approach in terms of F_1 measure is higher than for a state-of-the-art tagger based on conditional random fields (CRF), a freely available dictionary-based approach and an abbreviation resolver. We evaluate different parameters and their impact in our filtering approach and discuss the results. Note that this approach does not take any information about the specific abbreviation into account and can therefore be expected to generalize to names not considered in our corpus.

The main contributions of this paper are:

(i) We consider the problem of recognizing highly ambiguous abbreviations that frequently do not denote proteins as a task that has so far attracted only limited attention.

(ii) We show that the recognition of such ambiguous mentions is important as their string representation is frequent in collections such as MEDLINE.

(iii) We show, however, that this set of ambiguous names is underrepresented in corpora commonly used for system design and development. Such corpora do not provide a sufficient data basis for studying the phenomenon or for training systems that appropriately handle such ambiguous abbreviation. We contribute a manually annotated corpus of 2174 occurrences of ambiguous abbreviations.

(iv) We propose a filtering method for classifying ambiguous abbreviations as denoting a protein or not. We show that this method has a positive impact on the overall performance of named entity recognition systems.

2 Related Work

The task of gene/protein recognition consists in the classification of terms as actually denoting a gene/protein or not. The task is typically either tackled by using machine learning or dictionary-based approaches. Machine learning approaches rely on appropriate features describing the local context of the term to be classified and induce a model to perform the classification from training data. Conditional random fields have shown to yield very good results on the task (Klinger et al., 2007; Leaman and Gonzalez, 2008; Kuo et al., 2007; Settles, 2005).

Dictionary-based approaches rely on an explicit dictionary of gene/protein names that are matched in text. Such systems are common in practice due to the low overhead required to adapt and maintain the system, essentially only requiring to extend the dictionary. Examples of commercial systems are ProMiner (Fluck et al., 2007) or I2E (Bandy et al., 2009); a popular free system is made available by Hakenberg et al. (2011).

Such dictionary-based systems typically incorporate rules for filtering false positives. For instance, in ProMiner (Hanisch et al., 2003), ambiguous synonyms are only accepted based on external dictionaries and matches in the context. Abbreviations are only accepted if a long form matches all parts of the abbreviation in the context (following Schwartz and Hearst (2003)). Similarly, Hakenberg et al. (2008) discuss global disambiguation on the document level, such that all mentions of a string in one abstract are uniformly accepted as denoting an entity or not.

A slightly different approach is taken by the web-service GeneE[5] (Schuemie et al., 2010): Entering a query as a gene/protein in the search field generates

[5] http://biosemantics.org/geneE

	MEDLINE		BioCreative2		GENIA	
Protein	# Tokens	% tagged	# Tokens	% of genes	# Tokens	% of genes
SAH	30019	6.1 %	2	0 %	0	
MOX	16007	13.1 %	0		0	
PLS	11918	25.9 %	0		0	
CLU	1077	29.1 %	0		0	
CLI	1957	4.8 %	4	0 %	0	
HF	42563	7.9 %	8	62.5 %	4	0 %
AHR	21525	75.7 %	12	91.7 %	0	
COPD	44125	0.6 %	6	0 %	0	

Table 2: Coverage of ambiguous abbreviations in MEDLINE, BioCreative2 and GENIA corpora. The percentage of tokens tagged as a gene/protein in MEDLINE (% tagged) is determined with a conditional random field in the configuration described by Klinger et al. (2007), but without dictionary-based features to foster the usage of contextual features). The percentages of genes/proteins (% of genes) in BC2 and GENIA are based on the annotations in these corpora.

a query to *e. g.* PubMed®[6] with the goal to limit the number of false positives.

Previous to the common application of CRFs, other machine learning methods have been popular as well for the task of entity recognition. For instance, Mitsumori et al. (2005) and Bickel et al. (2004) use a support vector machine (SVM) with part-of-speech information and dictionary-based features, amongst others. Zhou et al. (2005) use an ensemble of different classifiers for recognition.

In contrast to this application of a classifier to solve the recognition task entirely, other approaches (including the one in this paper) aim at filtering specifically ambiguous entities from a previously defined set of challenging terms. For instance, Al-mubaid (2006) utilize a word-based classifier and a mutual information-based feature selection to achieve a highly discriminating list of terms which is applied for filtering candidates.

Similarly to our approach, Tsuruoka and Tsujii (2003) use a classifier, in their case a naïve Bayes approach, to learn which entities to filter from the candidates generated by a dictionary-based approach. They use word based features in the context including the candidate itself. Therefore, the approach is focused on specific entities.

Gaudan et al. (2005) use an SVM and a dictionary of long forms of abbreviations to assign them a specific meaning, taking contextual information into account. However, their machine learning approach is trained on each possible sense of an abbreviation. In contrast, our approach consists in deciding if a term is used as a protein or not. Further, we do not train to detect specific, previously given senses.

Xu et al. (2007) apply text similarity measures to decide about specific meanings of mentions. They focus on the disambiguation between different entities. A corpus for word sense disambiguation is automatically built based on MeSH annotations by Jimeno-Yepes et al. (2011). Okazaki et al. (2010) build a sense inventory by automatically applying patterns on MEDLINE and use this in a logistic regression approach.

Approaches are typically evaluated on freely available resources like the BioCreative Gene Mention Task Corpus, to which we refer as BC2 (Smith et al., 2008), or the GENIA Corpus (Kim et al., 2003). When it comes to identifying particular proteins by linking the protein in question to some protein in an external database – a task we do not address in this paper – the BioCreative Gene Normalization Task Corpus is a common resource (Morgan et al., 2008).

In contrast to these previous approaches, our method is not tailored to a particular set of entities or meanings, as the training methodology abstracts from specific entities. The model, in fact, knows nothing about the abbreviations to be classified and does not use their surface form as a feature, such that it can be applied to any unseen gene/protein term. This leads to a simpler model that is applicable to a wide range of gene/protein term candidates. Our cross-entity evaluation regime clearly corroborates this.

3 Data

We focus on eight ambiguous abbreviations of gene/protein names. As shown in Table 2, these homonyms occur relatively frequently in MEDLINE but are underrepresented in the BioCreative 2 entity

[6] `http://www.ncbi.nlm.nih.gov/pubmed/`

Protein	Pos. Inst.	Neg. Inst.	Total
SAH	5	349	354
MOX	62	221	283
PLS	1	206	207
CLU	235	30	265
CLI	11	211	222
HF	2	353	355
AHR	53	80	133
COPD	0	250	250

Table 3: Number of instances per protein in the annotated data set and their positive/negative distribution

recognition data set and the GENIA corpus which are both commonly used for developing and evaluating gene recognition approaches. We compiled a corpus from MEDLINE by randomly sampling 100 abstracts for each of the eight abbreviations (81 for MOX) such that each abstract contains at least one mention of the respective abbreviation. One of the authors manually annotated the mentions of the eight abbreviations under consideration to be a gene/protein entity or not. These annotations were validated by another author. Both annotators disagreed in only 2% of the cases. The numbers of annotations, including their distribution over positive and negative instances, are summarized in Table 3. The corpus is made publicly available at `http://dx.doi.org/10.4119/unibi/2673424` (Hartung and Zwick, 2014).

In order to alleviate the imbalance of positive and negative examples in the data, additional positive examples have been gathered by manually searching PubMed[7]. At this point, special attention has been paid to extract only instances denoting the correct gene/protein corresponding to the full long name, as we are interested in assessing the impact of examples of a particularly high quality. This process yields 69 additional instances for AHR (distributed over 11 abstracts), 7 instances (3 abstracts) for HF, 14 instances (2 abstracts) for PLS and 15 instances (7 abstracts) for SAH. For the other gene/proteins in our dataset, no additional positive instances of this kind could be retrieved using PubMed. In the following, this process will be referred to as *manual instance generation*. This additional data is used for training only.

4 Gene Recognition by Filtering

We frame gene/protein recognition from ambiguous abbreviations as a filtering task in which a set of candidate tokens is classified into entities and non-entities. In this paper, we assume the candidates to be generated by a simple dictionary-based approach taking into account all tokens that match the abbreviation under consideration.

4.1 Filtering Strategies

We consider the following filtering approaches:

- *SVM* classifies the occurring terms based on a binary support vector machine.
- *CRF* classifies the occurring terms based on a conditional random field (configured as described by Klinger et al. (2007)) trained on the concatenation of BC2 data and our newly generated corpus. This setting thus corresponds to state-of-the-art performance on the task.
- *CRF∩SVM* considers the candidate an entity if both the standard CRF and the SVM from the previous steps yield a positive prediction.
- *HRCRF∩SVM* is the same as the previous step, but the output of the CRF is optimized towards high recall by joining the recognition of entities of the five most likely Viterbi paths.
- *CRF→SVM* is similar to the first setting, but the output of the CRF is taken into account as a feature in the SVM.

4.2 Features for Classification

Our classifier uses local contextual and global features. Local features focus on the immediate context of an instance, whereas global features encode abstract-level information. Throughout the following discussion, t_i denotes a token at position i that corresponds to a particular abbreviation to be classified in an abstract A. Note that we blind the actual representation of the entity to be able to generalize to all genes/proteins, not being limited to the ones contained in our corpus.

4.2.1 Local Information

The feature templates *context-left* and *context-right* collect the tokens immediately surrounding an abbreviation in a window of size 6 (left) and 4 (right) in a bag-of-words-like feature generation. Additionally, the two tokens from the immediate context on each side are combined into bigrams.

The template *abbreviation* generates features if t_i occurs in brackets. It takes into account the minimal Levenshtein distance (*ld*, Levenshtein (1966))

[7] `http://www.ncbi.nlm.nih.gov/pubmed`

between all long forms L of the abbreviation (as retrieved from EntrezGene) in comparison to each string on the left of t_i (up to a length of seven, denoted by $t_{k:i}$ as the concatenation of tokens $t_k, \ldots, t_i$). Therefore, the similarity value $\mathit{sim}(t_i)$ taken into account is given by

$$\mathit{sim}(t_i) = \max_{l\in L; k\in[1:7]} 1 - \frac{\mathit{ld}(t_{k:i-1}, l)}{\max(|t_i|, |l|)},$$

where the denominator is a normalization term. The features used are generated by cumulative binning of $\mathit{sim}(t_i)$.

The feature *tagger*$_{\mathit{local}}$ takes the prediction of the CRF for t_i into account. Note that this feature is only used in the CRF→SVM setting.

4.2.2 Global Information

The feature template *unigrams* considers each word in A as a feature. There is no normalization or frequency weighting. Stopwords are ignored[8]. Occurrences of the same string as t_i are blinded.

The feature *tagger*$_{\mathit{global}}$ collects all tokens in A other than t_i that are tagged as an entity by the CRF. In addition, the cardinality of these entities in A is taken into account by cumulative binning.

The feature *long form* holds if one of the long forms previously defined to correspond with the abbreviation occurs in the text (in arbitrary position).

Besides using all features, we perform a greedy search for the best feature set by wrapping the best model configuration. A detailed discussion of the feature selection process follows in Section 5.3.

4.2.3 Feature Propagation

Inspired by the “one sense per discourse” heuristic commonly adopted in word sense disambiguation (Gale et al., 1992), we apply two feature combination strategies. In the following, n denotes the number of occurrences of the abbreviation in an abstract.

In the setting *propagation*$_{\mathit{all}}$, $n - 1$ identical *linked instances* are added for each occurrence. Each new instance consists of the disjunction of the feature vectors of all occurrences. Based on the intuition that the first mention of an abbreviation might carry particularly valuable information, *propagation*$_{\mathit{first}}$ introduces one additional linked instance for each occurrence, in which the feature vector is joined with the first occurrence.

Setting	P	R	F$_1$
SVM	0.81	0.45	0.58
CRF∩SVM	**0.99**	0.26	0.41
HRCRF∩SVM	0.95	0.27	0.42
CRF→SVM	0.83	0.49	0.62
CRF→SVM+FS	0.97	**0.74**	**0.84**
GNAT	0.73	0.45	0.56
CRF	0.55	0.43	0.48
AcroTagger	0.92	0.63	0.75
Long form	0.98	0.65	0.78
lex	0.18	1.00	0.32

Table 4: Overall micro-averaged results over eight genes/proteins. For comparison, we show the results of a default run of GNAT (Hakenberg et al., 2011), a CRF trained on BC2 data (Klinger et al., 2007), AcroTagger (Gaudan et al., 2005), and a simple approach of accepting every token of the respective string as a gene/protein entity (lex). Feature selection is denoted with +FS.

In both settings, all original and linked instances are used for training, while during testing, original instances are classified by majority voting on their linked instances. For *propagation*$_{\mathit{all}}$, this results in classifying each occurrence identically.

5 Experimental Evaluation

5.1 Experimental Setting

We perform a cross-entity evaluation, in which we train the support vector machine (SVM) on the abstracts of 7 genes/proteins from our corpus and test on the abstracts for the remaining entities, *i. e.*, the model is evaluated only on tokens representing entities which have never been seen labeled during training. The CRFs are trained analogously with the difference that the respective set used for training is augmented with the BioCreative 2 Training data. The average numbers of precision, recall and F$_1$ measure are reported.

As a baseline, we report the results of a simple lexicon-based approach assuming that all tokens denote an entity in all their occurrences (lex). In addition, the baseline of accepting an abbreviation as gene/protein if the long form occurs in the same abstract is reported (Long form). Moreover, we compare our results with the publicly available toolkit GNAT (Hakenberg et al., 2011)[9] and the CRF ap-

[8] Using the stopword list at `http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/`, last accessed on March 25, 2014

[9] The gene normalization functionality of GNAT is not taken into account here. We acknowledge that this comparison

proach as described in Section 4. In addition, we take into account the AcroTagger[10] that resolves abbreviations to their most likely long form which we manually map to denoting a gene/protein or not.

5.2 Results

5.2.1 Overall results

In Table 4, we summarize the results of the recognition strategies introduced in Section 4. The lexical baseline clearly proves that a simple approach without any filtering is not practical. GNAT adapts well to ambiguous short names and turns out as a competitive baseline, achieving an average precision of 0.73. In contrast, the filtering capacity of a standard CRF is, at best, mediocre. The long form baseline is very competitive with an F_1 measure of 0.78 and a close-to-perfect precision. The results of AcroTagger are similar to this long form baseline.

We observe that the SVM outperforms the CRF in terms of precision and recall (by 10 percentage points in F_1). Despite not being fully satisfactory either, these results indicate that global features which are not implemented in the CRF are of importance. This is confirmed by the CRF∩SVM setting, where CRF and SVM are stacked: This filtering procedure achieves the best precision across all models and baselines, whereas the recall is still limited. Despite being designed for exactly this purpose, the HRCRF∩SVM combination can only marginally alleviate this problem, and only at the expense of a drop in precision.

The best trade-off between precision and recall is offered by the CRF→SVM combination. This setting is not only superior to all other variants of combining a CRF with an SVM, but outperforms GNAT by 6 points in F_1 score, while being inferior to the long form baseline. However, performing feature selection on this best model using a wrapper approach (CRF→SVM+FS) leads to the overall best result of $F_1 = 0.84$, outperforming all other approaches and all baselines.

5.2.2 Individual results

Table 5 summarizes the performance of all filtering strategies broken down into individual entities. Best results are achieved for AHR, MOX and CLU. COPD forms a special case as no examples for the occurrence as a gene/protein are in the data; however the results show that the system can handle such a special distribution.

SVM and CRF are mostly outperformed by a combination of both strategies (except for CLI and HF), which shows that local and global features are highly complementary in general. Complementary cases generally favor the CRF→SVM strategy, except for PLS, where stacking is more effective.

In SAH, the pure CRF model is superior to all combinations of CRF and SVM. Apparently, the global information as contributed by the SVM is less effective than local contextual features as available to the CRF in these cases. In SAH and CLI, moreover, the best performance is obtained by the AcroTagger.

5.2.3 Impact of instance generation

All results reported in Tables 4 and 5 refer to configurations in which additional training instances have been created by manual instance generation. The impact of this method is analyzed in Table 6. The first column reports the performance of our models on the randomly sampled training data. In order to obtain the results in the second column, manual instance generation has been applied.

The results show that all our recognition models generally benefit from additional information that helps to overcome the skewed class distribution of the training data. Despite their relatively small quantity and uneven distribution across the gene/protein classes, including additional external instances yields a strong boost in all models. The largest difference is observed in SVM ($\Delta F_1 = +0.2$) and CRF→SVM ($\Delta F_1 = +0.16$). Importantly, these improvements include both precision and recall.

5.3 Feature Selection

The best feature set (*cf.* CRF→SVM+FS in Table 4) is determined by a greedy search using a wrapper approach on the best model configuration CRF→SVM. The results are depicted in Table 7. In each iteration, the table shows the best feature set detected in the previous iteration and the results for each individual feature when being added to this set. In each step, the best individual feature is kept for the next iteration. The feature analysis starts from the *long form* feature as strong baseline. The added features are, in that order, *context*, *tagger*$_{global}$, and *propagation*$_{all}$.

Overall, feature selection yields a considerable

might be seen as slightly inappropriate as the focus of GNAT is different.

[10] `ftp://ftp.ebi.ac.uk/pub/software/textmining/abbreviation_resolution/`, accessed April 23, 2014

	AHR			CLI			CLU			COPD		
Setting	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
SVM	1.00	0.72	0.84	0.30	0.27	0.29	1.00	0.41	0.58	0.00	1.00	0.00
CRF∩SVM	1.00	0.70	0.82	0.00	0.00	0.00	1.00	0.15	0.26	1.00	1.00	1.00
HRCRF∩SVM	1.00	0.70	0.82	1.00	0.00	0.00	1.00	0.16	0.28	1.00	1.00	1.00
CRF→SVM	0.96	0.83	0.89	0.30	0.27	0.29	1.00	0.40	0.57	0.00	1.00	0.00
CRF→SVM+FS	0.93	0.98	0.95	0.50	0.09	0.15	0.99	0.84	0.91	1.00	1.00	1.00
GNAT	0.74	0.66	0.70	1.00	0.18	0.31	0.97	0.52	0.68	1.00	1.00	1.00
CRF	0.52	0.98	0.68	0.00	0.00	0.00	1.00	0.20	0.33	0.00	1.00	0.00
AcroTagger	1.00	0.60	0.75	1.00	0.82	0.90	1.00	0.00	0.00	1.00	1.00	1.00
Long form	1.00	0.96	0.98	1.00	0.09	0.17	0.99	0.80	0.88	1.00	1.00	1.00
lex	0.40	1.00	0.57	0.05	1.00	0.09	0.89	1.00	0.94	0.00	1.00	0.00

	HF			MOX			PLS			SAH		
Setting	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
SVM	0.25	1.00	0.40	0.87	0.44	0.58	0.14	1.00	0.25	0.00	0.00	0.00
CRF∩SVM	1.00	0.00	0.00	1.00	0.39	0.56	1.00	1.00	1.00	1.00	0.00	0.00
HRCRF∩SVM	1.00	0.00	0.00	1.00	0.39	0.56	0.20	1.00	0.33	1.00	0.00	0.00
CRF→SVM	0.25	1.00	0.40	0.91	0.63	0.74	0.50	1.00	0.67	1.00	0.00	0.00
CRF→SVM+FS	1.00	0.00	0.00	1.00	0.37	0.54	0.00	0.00	0.00	1.00	0.00	0.00
GNAT	1.00	0.00	0.00	0.38	0.08	0.14	0.00	0.00	0.00	0.00	0.00	0.0
CRF	0.00	0.00	0.00	0.43	0.90	0.59	0.14	1.00	0.25	1.00	0.50	0.67
AcroTagger	0.33	1.00	0.50	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.60	0.75
Long form	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
lex	0.01	1.00	0.02	0.22	1.00	0.36	0.00	1.00	0.01	0.01	1.00	0.03

Table 5: Results for the eight genes/proteins and results for our different recognition schemes.

	randomly sampled			+instance generation		
	P	R	F_1	ΔP	ΔR	ΔF_1
SVM	0.73	0.25	0.38	+0.08	+0.20	+0.20
CRF∩SVM	1.00	0.17	0.29	-0.01	+0.09	+0.13
HRCRF∩SVM	0.97	0.18	0.30	-0.02	+0.09	+0.12
CRF→SVM	0.79	0.32	0.46	+0.05	+0.17	+0.16
CRF→SVM+FS	0.99	0.60	0.75	-0.02	+0.14	+0.09

Table 6: Impact of increasing the randomly sampled training set by adding manually curated additional positive instances (+instance generation), measured in terms of the increase in precision, recall and F_1 (ΔP, ΔR, ΔF_1).

boost in recall, while precision remains almost constant. Surprisingly, the *unigrams* feature has a particularly strong negative impact on overall performance.

While the global information contributed by the CRF turns out very valuable, accounting for most of the improvement in recall, local tagger information is widely superseded by other features. Likewise, the *abbreviation* feature does not provide any added value to the model beyond what is known from the *long form* feature.

Comparing the different feature propagation strategies, we observe that *propagation*$_{all}$ outperforms *propagation*$_{first}$.

5.4 Discussion

Our experiments show that the phenomena investigated pose a challenge to all gene recognition paradigms currently available in the literature, *i. e.*, dictionary-based, machine-learning-based (*e. g.* using a CRF), and classification-based filtering.

Our results indicate that stacking different methods suffers from a low recall in early steps of the workflow. Instead, a greedy approach that considers all occurrences of an abbreviation as input to a filtering approach yields the best performance. Incorporating information from a CRF as features into a SVM outperforms all baselines at very high levels of precision; however, the recall still leaves room for improvement.

Iter.	Feature Set	P	R	F$_1$	ΔF$_1$
1	***long form***	**0.98**	**0.65**	**0.78**	
	+propagation$_{1^{st}}$	0.98	0.65	0.78	+0.00
	+propagation$_{all}$	0.98	0.65	0.78	+0.00
	+tagger$_{local}$	0.72	0.81	0.76	-0.02
	+tagger$_{global}$	0.55	0.79	0.65	-0.13
	+context	0.98	0.67	0.79	**+0.01**
	+abbreviation	0.98	0.65	0.78	+0.00
	+unigrams	0.71	0.43	0.53	-0.25
2	***long form*** ***+context***	**0.98**	**0.67**	**0.79**	
	+propagation$_{1^{st}}$	0.98	0.67	0.79	+0.00
	+propagation$_{all}$	0.96	0.70	0.81	+0.02
	+tagger$_{local}$	0.98	0.70	0.82	+0.03
	+tagger$_{\mathbf{global}}$	0.97	0.72	0.83	**+0.04**
	+abbreviation	0.98	0.67	0.80	+0.01
	+unigrams	0.77	0.39	0.52	-0.27
3	***long form*** ***+context*** ***+tagger***$_{\mathbf{global}}$	**0.97**	**0.72**	**0.83**	
	+propagation$_{1^{st}}$	0.97	0.71	0.82	-0.01
	+propagation$_{\mathbf{all}}$	0.97	0.74	0.84	**+0.01**
	+tagger$_{local}$	0.97	0.72	0.82	-0.01
	+abbreviation	0.97	0.72	0.82	-0.01
	+unigrams	0.77	0.44	0.56	-0.27
4	***long form*** ***+context*** ***+tagger***$_{\mathbf{global}}$ ***+propagation***$_{\mathbf{all}}$	**0.97**	**0.74**	**0.84**	
	+tagger$_{local}$	0.90	0.66	0.76	-0.08
	+abbreviation	0.97	0.74	0.84	-0.00
	+unigrams	0.80	0.49	0.61	-0.23

Table 7: Greedy search for best feature combination in CRF→SVM (incl. additional positives).

In a feature selection study, we were able to show a largely positive overall impact of features that extend local contextual information as commonly applied by state-of-the-art CRF approaches. This ranges from larger context windows for collecting contextual information over abstract-level features to feature propagation strategies. However, feature selection is not equally effective in all individual classes (*cf.* Table 5).

The benefits due to feature propagation indicate that several instances of the same abbreviation in one abstract should not be considered independently of one another, although we could not verify the intuition that the first mention of an abbreviation introduces particularly valuable information for classification.

Overall, our results seem encouraging as the machinery and the features used are in general successful in determining whether an abbreviation actually denotes a gene/protein or not. The best precision/recall balance is obtained by adding CRF information as features into the classifier.

As we have shown in the cross-entity experiment setting, the system is capable of generalizing to other unseen entities. For a productive system, we assume our workflow to be applied to specific abbreviations such that the performance on other entities (and therefore on other corpora) is not substantially influenced.

6 Conclusions and Outlook

The work reported in this paper was motivated from the practical need for an effective filtering method for recognizing genes/proteins from highly ambiguous abbreviations. To the best of our knowledge, this is the first approach to tackle gene/protein recognition from ambiguous abbreviations in a systematic manner without being specific for the particular instances of ambiguous gene/protein homonyms considered.

The proposed method has been proven to allow for an improvement in recognition performance when added to an existing NER workflow. Despite being restricted to eight entities so far, our approach has been evaluated in a strict cross-entity manner, which suggests sufficient generalization power to be extended to other genes as well.

In future work, we plan to extend the data set to prove the generalizability on a larger scale and on an independent test set. Furthermore, an inclusion of the features presented in this paper into the CRF will be evaluated. Moreover, assessing the impact of the global features that turned out beneficial in this paper on other gene/protein inventories seems an interesting path to explore. Finally, we will investigate the prospects of our approach in an actual black-box evaluation setting for information retrieval.

Acknowledgements

Roman Klinger has been funded by the “It’s OWL” project (“Intelligent Technical Systems Ostwestfalen-Lippe”, `http://www.its-owl.de/`), a leading-edge cluster of the German Ministry of Education and Research. We thank Jörg Hakenberg and Philippe Thomas for their support in performing the baseline results with GNAT. Additionally, we thank the reviewers of this paper for their very helpful comments.

References

Hisham Al-mubaid. 2006. Biomedical term disambiguation: An application to gene-protein name disambiguation. In *In IEEE Proceedings of ITNG06*.

Judith Bandy, David Milward, and Sarah McQuay. 2009. Mining protein-protein interactions from published literature using linguamatics i2e. *Methods Mol Biol*, 563:3–13.

Steffen Bickel, Ulf Brefeld, Lukas Faulstich, Jörg Hakenberg, Ulf Leser, Conrad Plake, and Tobias Scheffer. 2004. A support vector machine classifier for gene name recognition. In *In Proceedings of the EMBO Workshop: A Critical Assessment of Text Mining Methods in Molecular Biology*.

Juliane Fluck, Heinz Theodor Mevissen, Marius Oster, and Martin Hofmann-Apitius. 2007. ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 149–151, Madrid, Spain.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pages 233–237, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sylvain Gaudan, Harald Kirsch, and Dietrich Rebholz-Schuhmann. 2005. Resolving abbreviations to their senses in medline. *Bioinformatics*, 21(18):3658–3664.

Jörg Hakenberg, Conrad Plake, Robert Leaman, Michael Schroeder, and Graciela Gonzalez. 2008. Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, 24(16):i126–i132, Aug.

Jörg Hakenberg, Martin Gerner, Maximilian Haeussler, Ills Solt, Conrad Plake, Michael Schroeder, Graciela Gonzalez, Goran Nenadic, and Casey M. Bergman. 2011. The GNAT library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771, Oct.

Daniel Hanisch, Juliane Fluck, Heinz-Theodor Mevissen, and Ralf Zimmer. 2003. Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput*, pages 403–414.

Matthias Hartung and Matthias Zwick. 2014. A corpus for the development of gene/protein recognition from rare and ambiguous abbreviations. Bielefeld University. doi:10.4119/unibi/2673424.

Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223.

J-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. Genia corpus–semantically annotated corpus for biotextmining. *Bioinformatics*, 19 Suppl 1:i180–i182.

Roman Klinger, Christoph M. Friedrich, Juliane Fluck, and Martin Hofmann-Apitius. 2007. Named Entity Recognition with Combinations of Conditional Random Fields. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, Madrid, Spain, April.

Cheng-Ju Kuo, Yu-Ming Chang, Han-Shen Huang, Kuan-Ting Lin, Bo-Hou Yang, Yu-Shi Lin, Chun-Nan Hsu, and I-Fang Chung. 2007. Rich feature set, unication of bidirectional parsing and dictionary filtering for high f-score gene mention tagging. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, Madrid, Spain, April.

Robert Leaman and Graciela Gonzalez. 2008. Banner: An executable survey of advances in biomedical named entity recognition. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Tiffany Murray, and Teri E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 652–663. World Scientific.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi, and Hirohumi Doi. 2005. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6 Suppl 1:S8.

Alexander A. Morgan, Zhiyong Lu, Xinglong Wang, Aaron M. Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jrg Hakenberg, Chengjie Sun, Heng-hui Liu, Rafael Torres, Michael Krauthammer, William W. Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K Bretonnel Cohen, and Lynette Hirschman. 2008. Overview of biocreative ii gene normalization. *Genome Biol*, 9 Suppl 2:S3.

Naoaki Okazaki, Sophia Ananiadou, and Jun'ichi Tsujii. 2010. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9):1246–1253, May.

Martijn J. Schuemie, Ning Kang, Maarten L. Hekkelman, and Jan A. Kors. 2010. Genee: gene and protein query expansion with disambiguation. *Bioinformatics*, 26(1):147–148, Jan.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput*, pages 451–462.

Burr Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, Jul.

Larry Smith, Lorraine K. Tanabe, Rie Johnson nee J. Ando, Cheng-Ju J. Kuo, I-Fang F. Chung, Chun-Nan N. Hsu, Yu-Shi S. Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han T. Tsai, Hong-Jie J. Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña López, Jacinto Mata, and W. John Wilbur. 2008. Overview of BioCreative II gene mention recognition. *Genome biology*, 9 Suppl 2(Suppl 2):S2+.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2003. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 41–48, Sapporo, Japan, July. Association for Computational Linguistics.

Hua Xu, Jung-Wei Fan, George Hripcsak, Eneida A Mendonça, Marianthi Markatou, and Carol Friedman. 2007. Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics*, 23(8):1015–1022.

GuoDong Zhou, Dan Shen, Jie Zhang, Jian Su, and SoonHeng Tan. 2005. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*, 6 Suppl 1:S7.