Recognising Nested Named Entities in Biomedical Text

Beatrice Alex, Barry Haddow and Claire Grover
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW, UK
{balex,bhaddow,grover}@inf.ed.ac.uk

Abstract

Although recent named entity (NE) annotation efforts involve the markup of nested entities, there has been limited focus on recognising such nested structures. This paper introduces and compares three techniques for modelling and recognising nested entities by means of a conventional sequence tagger. The methods are tested and evaluated on two biomedical data sets that contain entity nesting. All methods yield an improvement over the baseline tagger that is only trained on flat annotation.

1 Introduction

Traditionally, named entity recognition (NER) has focussed on entities which are *continuous*, *non-nested* and *non-overlapping*. In other words, each token in the text belongs to at most one entity, and NEs consist of a continuous sequence of tokens. However, in some situations, it may make sense to relax these restrictions, for example by allowing entities to be *nested* inside other entities, or allowing *discontinuous* entities. GENIA (Ohta et al., 2002) and BioInfer (Pyysalo et al., 2007) are examples of recently produced NE-annotated biomedical corpora where entities nest. Corpora in other domains, for example the ACE[1] data, also contain nested entities.

This paper compares techniques for recognising nested entities in biomedical text. The difficulty of this task is that the standard method for converting NER to a sequence tagging problem with BIO-encoding (Ramshaw and Marcus, 1995), where each token is assigned a tag to indicate whether it is at the beginning (B), inside (I), or outside (O) of an entity, is not directly applicable when tokens belong to more than one entity. Here we explore methods of reducing the nested NER problem to one or more BIO problems so that existing NER tools can be used.

This paper is organised as follows. In Section 2, the problem of nested entities is introduced and motivated with examples from GENIA and our EPPI (enriched protein-protein interaction) data. Related work is reviewed in Section 3. The proposed techniques enabling NER for nested NEs are explained in Section 4. Section 5 details the experimental setup, including descriptive statistics of the corpora and specifics of the classifier. The results of comparing different tagging methods are analysed in Section 6, with a discussion and conclusion in Section 7.

2 Nested Entities

The majority of previous work on NER is conducted using data sets annotated either with continuous, non-nested and non-overlapping NEs or an annotation scheme reduced to a flat annotation of a similar kind in order to simplify the recognition task. However, annotated corpora often contain entities that are nested or discontinuous. For example, the GENIA corpus contains nested entities such as:

<RNA><DNA>CIITA</DNA> mRNA</RNA>

where the string “CIITA” denotes a DNA and the entire string “CIITA mRNA” refers to an RNA. Such nesting complicates the task of traditional NER systems, which generally rely on data represented with the BIO encoding or other flat annotation variations thereof. The majority of NER studies on corpora

[1] http://www.nist.gov/speech/tests/ace/index.htm

BioNLP 2007: Biological, translational, and clinical language processing, pages 65–72,
Prague, June 2007.

GENIA		EPPI	
Count	Nesting	Count	Nesting
3,614	(other_name (protein t) t)	1,698	(fusion (protein t) t (protein t))
907	(DNA (protein t) t)	1,269	(drug/compound (protein t))
856	(protein (protein t) t)	455	(fusion (fragment t) t (protein t))
661	(protein t (protein t))	412	(protein (protein t) t)
546	(other_name (DNA t) t)	361	(complex (protein t) t (protein t))
541	(other_name t (other_name t))	298	(fusion (protein t) t (fragment t))
470	(cell_type t (cell_type t))	246	(fragment t (fragment t))
351	(DNA t (DNA t))	241	(cell_line t (cell_line t))
326	(other_name (virus t) t)	207	(fragment (protein t))
262	(other_name (lipid t) t)	201	(fusion (protein t) t (mutant t))

Table 1: 10 most frequent types of nesting in the GENIA corpus and the combined TRAIN and DEVTEST sections of the EPPI data (see Section 5.1), where t represents the text.

containing nested structures focus on recognising the outermost (non-embedded) entities (e.g. Kim et al. 2004) , as they contain the most information, including that of embedded entities (Zhang et al., 2004). This enables a simplification of the recognition task to a sequential analysis problem.

Our aim is to recognise all levels of NE nesting occurring in two biomedical corpora: the GENIA corpus (Version 3.02) and the EPPI corpus (see Section 5.1). The latter data set was collected and annotated as part of the TXM project. Its annotation contains 9 different biomedical entities. While the GENIA corpus contains nested entities up to a level of four layers of embedding, the nested entities in the EPPI corpus only have three layers. Table 1 lists the ten most frequent types of entity nesting occurring in both corpora. In the remainder of the paper, we differentiate between:

embedded NEs: contained in other NEs

non-embedded NEs: not contained in other NEs

containing NEs: containing other NEs

non-containing NEs: not containing other NEs

The GENIA corpus is made up of a larger percentage of both embedded entity (18.61%) and containing entity (16.95%) mentions than the EPPI data (12.02% and 8.27%, respectively). In both corpora, nesting can occur in three different ways:

1. *Entities containing one or more shorter embedded entities.* Such nesting is very frequent in both data sets. For example, the DNA "IL-2 promoter" in the GENIA corpus contains the protein "IL-2". In the EPPI corpus, fusions and complexes often contain nested proteins, e.g. the complex "CBP/p300", where "CBP" and "p300" are marked as proteins.

2. *Entities with more than one entity type.* Although they occur in both data sets, they are very rare in the GENIA corpus. For example, the string "p21ras" is annotated both as DNA and protein. In the EPPI data, proteins can also be annotated as drug/compound, where it can be clearly established that the protein is used as a drug to affect the function of an organism, cell, or biological process.

3. *Coordinated entities.* Coordinated NEs account for approximately 2% of all NEs in the GENIA and EPPI data. In the original corpora they are annotated differently, but for this work they are all converted to a common format.[2] The outermost annotation of coordinated structures and any continuous entity mark-up within them is retained. For example, in "human interleukin-2 and -4" both the continuous embedded entity "human interleukin-2" and the entire string are marked as proteins. The markup for discontinuous embedded entities, like "human interleukin-4" in the previous example, is not retained, as they could be derived in a post-processing step once nested entities are recognised.

3 Related Work

In previous work addressing nested entities, Shen et al. (2003), Zhang et al. (2004), Zhou et al. (2004), Zhou (2006), and Gu (2006) considered the GENIA

[2]Both corpora are represented in XML with standoff annotation, potentially allowing overlapping NEs.

corpus, where nested entities are relatively frequent. All these studies ignore embedded entities occurring in coordinated structures and only retain their outermost annotation. Shen et al. (2003), Zhang et al. (2004), and Zhou et al. (2004) all report on a rule-based approach to dealing with nested NEs in the GENIA corpus (Version 3.0) in combination with a Hidden Markov Model (HMM) that first recognises innermost NEs. They use four basic hand-crafted patterns and a combination thereof to generate nesting rules from the training data and thereby derive NEs containing the innermost NEs. The experimental setup of these studies differs slightly. While Shen et al. (2003) and Zhang et al. (2004) report results testing on 4% of the abstracts in the GENIA corpus, Zhou et al. (2004) report 10-fold cross-validation scores. Zhou (2006) applies the same rule-based method for dealing with nested entities to the output of a mutual information independence model (MIIM) combined with a support vector machine (SVM) plus sigmoid. His results are based on 5-fold cross-validation on the GENIA corpus (Version 3.0). In each of the studies, the rule-based approach to nested entities results in an improvement of between 3.0 and 3.5 points in $F1$ over the baseline model. However, as explicitly stated by Shen et al. (2003) and Zhang et al. (2004), this evaluation is limited to non-embedded (i.e. top-level and non-nested) entities. The highest overall $F1$-score reported for all entities in the GENIA corpus is 71.2 (Zhou, 2006), which again only appears to reflect the performance on non-embedded entities.

Zhang et al. (2004) also compare the rule-based method with HMM-based cascaded recognition that extends iteratively from the shortest to the longest entities. Their basic HMM model is combined with HMM models trained on transformed cascaded annotations. During training, embedded entity terms are replaced by their entity type as a way of unnesting the data. During testing, subsequent iterations rely on the tagging of the first recognition pass and are repeated until no more entities are recognised. However, this method only results in an improvement of 1.2 points in $F1$ over their basic classifier.

Gu (2006) reports results on recognising nested entities in the GENIA corpus (Version 3.02) when training an SVM-light binary classifier to recognise either proteins or DNA. Training with the outermost labelling yields better performance on recognising outermost entities and, conversely, using the inner labelling results in highest scores for recognising inner entities. The best exact match $F1$-scores of 73.0 and 47.5 for proteins and DNA, respectively, are obtained when training on data with inner labelling and evaluating on the inner entities.

McDonald et al. (2005) propose structured multilabel classification as opposed to sequential labelling for dealing with nested, discontinuous, and overlapping NEs. This approach uses a novel text segment representation in preference to the BIO-encoding. Their corpus contains MEDLINE abstracts on the inhibition of the enzyme CYP450 (Kulick et al., 2004), specifically those abstracts that contain at least one overlapping and one discontinuous annotation. While this data does not contain nested NEs, discontinuous and overlapping NEs make up 6% of all NEs. The classifier performs competitively with sequential tagging models on continuous and non-overlapping entities for NER and noun phrase chunking. On discontinuous and overlapping NEs in the biomedical data alone, its best performance is 56.25 $F1$. As the corpus does not contain nested NEs, it would be of interest to investigate the algorithm's performance on the GENIA corpus.

4 Modelling Techniques

As large amounts of time and effort have been devoted to work on non-nested NER using the BIO-encoding approach, it would be useful if this work could be easily applied to nested NER. In this paper, three different ways of addressing nested NER will be compared: *layering*, *cascading*, and *joined label tagging*. All techniques aim to reduce the nested NER problem to one or more BIO problems, so that existing NER tools can be used. Table 2 shows an example representation for each modelling technique of the following two non-nested and nested entity annotations found in a GENIA abstract:

<multi_cell>mice</multi_cell>...
<other_name><RNA><protein>tumor necrosis factor-alpha</protein> (<protein>TNF- alpha</protein>) messenger RNA</RNA> levels</other_name>

In layering, each level of nesting is modelled as a separate BIO problem. The output of models trained on individual layers is combined subsequent to tagging by taking the union. Layers can be created

Token	Inside-out layering			Outside-in layering		
Model	Layer 1	Layer 2	Layer 3	Layer 3	Layer 2	Layer 1
mice	B-multi_cell	O	O	B-multi_cell	O	O
...	...	...	...	...	...	...
tumor	B-protein	B-RNA	B-other_name	B-other_name	B-RNA	B-protein
necrosis	I-protein	I-RNA	I-other_name	I-other_name	I-RNA	I-protein
factor-alpha	I-protein	I-RNA	I-other_name	I-other_name	I-RNA	I-protein
(	O	I-RNA	I-other_name	I-other_name	I-RNA	O
TNF-alpha	B-protein	I-RNA	I-other_name	I-other_name	I-RNA	B-protein
)	O	I-RNA	I-other_name	I-other_name	I-RNA	O
messenger	O	I-RNA	I-other_name	I-other_name	I-RNA	O
RNA	O	I-RNA	I-other_name	I-other_name	I-RNA	O
levels	O	O	I-other_name	I-other_name	O	O

	Cascading			Joined label tagging
Model	All entity types	other	RNA	Joined labels
mice	B-multi_cell	O	O	B-multi_cell+O+O
...	...	...	...	...
tumor	B-protein	B-other_name	B-RNA	B-protein+B-RNA+B-other_name
necrosis	I-protein	I-other_name	I-RNA	I-protein+I-RNA+I-other_name
factor-alpha	I-protein	I-other_name	I-RNA	I-protein+I-RNA+I-other_name
(	O	I-other_name	I-RNA	O+I-RNA+I-other_name
TNF-alpha	B-protein	I-other_name	I-RNA	B-protein+I-RNA+I-other_name
)	O	I-other_name	I-RNA	O+I-RNA+I-other_name
messenger	O	I-other_name	I-RNA	O+I-RNA+I-other_name
RNA	O	I-other_name	I-RNA	O+I-RNA+I-other_name
levels	O	I-other_name	O	O+O+I-other_name

Table 2: Example representation of nested entities for various modelling techniques.

inside-out or *outside-in*. For inside-out layering, the first layer is made up of all non-containing entities, the second layer is composed of all those entities which only contain one layer of nesting, etc. Conversely, outside-in layering means that the first layer contains all non-embedded entities, the second layer contains all entities which are only contained within one outer entity, etc. Both directions of layering can be modelled using a conventional NE tagger.

Cascading reduces the nested NER task to several BIO problems by grouping one or more entity types and training a separate model for each group. Again, the output from individual models is combined during tagging. Subsequent models in the cascade may have access to the guesses of previous ones by means of a GUESS feature. The cascaded method is unable to recognise entities containing entities of the same type, which may be a drawback for some data sets. Cascading also raises the issue of how to group entity types. This is dependent on the types of entities that nest within a given data set and would potentially require large amounts of experimentation to determine the best combination. Moreover, training a model for each entity type lengthens training time considerably, and may degrade performance due to the dominance of the O tags for infrequent categories. It is possible, however, to create a cascaded tagger combining one model trained on all entity types with models trained on entity types that frequently contain other entities.

Finally, joined label tagging entails creating one tagging problem for all entities by concatenating the BIO tags of all levels of nesting. A conventional named entity recogniser is then trained on the data containing the joined labels. Once the classifier has assigned the joined labels during tagging, they are decoded into their original BIO format for each individual entity type. Compared to the other techniques, joined label tagging involves a much larger tag set, which can increase dramatically with the number of entity types occurring in a data set. This can result in data sparsity which may have a detrimental effect on performance.

5 Experimental Setup

5.1 Corpora

GENIA (V3.02), a large publicly available biomedical corpus annotated with biomedical NEs, is widely used in the text mining community (Cohen et al., 2005). This data set consists of 2,000 MEDLINE abstracts in the domain of molecular biology ($\simeq$0.5m tokens). The annotations used for the experiments reported here are based on the GENIA ontology, published in Ohta et al. (2002). It contains the following classes: amino acid monomer, atom, body part, carbohydrate, cell component, cell line, cell type, DNA, inorganic, lipid, mono-cell, multi-cell, nucleotide, other name, other artificial source, other organic compound, peptide, polynucleotide, protein, RNA, tissue, and virus. In this work, protein, DNA and RNA sub-types are collapsed to their super-type, as done in previous studies (e.g. Zhou 2006). To the best of our knowledge, no inter-annotator agreement (IAA) figures on the NE-annotation in the GENIA corpus are reported in the literature.

The EPPI corpus consists of 217 full-text papers selected from PubMed and PubMedCentral as containing protein-protein interactions (PPIs). The papers were either retrieved in XML or HTML, depending on availability, and converted to an internal XML format. Domain experts annotated all documents for NEs and PPIs, as well as extra (enriched) information associated with PPIs and normalisations of entities to publicly available ontologies. The entity annotations are the focus of the current work. The types of entities annotated in this data set are: complex, cell line, drug/compound, experimental method, fusion, fragment, modification, mutant, and protein. Out of the 217 papers, 125 were singly annotated, 65 were doubly annotated, and 27 were triply annotated. The IAA, measured by taking the $F1$ score of one annotator with respect to another when the same paper is annotated by two different annotators, ranges from 60.40 for the entity type mutant to 91.59 for protein, with an overall micro-averaged $F1$-score of 84.87. The EPPI corpus ($\simeq$2m tokens) is divided into three sections, TRAIN (66%), DEVTEST (17%), and TEST (17%), with TEST only to be used for final evaluation, and not to be consulted by the researchers in the development and feature optimisation phrase. The experiments described here involve the EPPI TRAIN and DEVTEST sets.

5.2 Pre-processing

All documents are passed through a sequence of pre-processing steps implemented using the LT-XML2 and LT-TTT2 tools (Grover et al., 2006) with the output of each step encoded in XML mark-up. Tokenisation and sentence splitting is followed by part-of-speech tagging with the Maximum Entropy Markov Model (MEMM) tagger developed by Curran and Clark (2003) (hereafter referred to as C&C) for the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003), trained on the MedPost data (Smith et al., 2004). Information on lemmatisation, as well as abbreviations and their long forms, is added using the *morpha* lemmatiser (Minnen et al., 2000) and the *ExtractAbbrev* script of Schwartz and Hearst (2003), respectively. A lookup step uses ontological information to identify scientific and common English names of species. Finally, a rule-based chunker marks up noun and verb groups and their heads (Grover and Tobin, 2006).

5.3 Named Entity Tagging

The C&C tagger, referred to earlier, forms the basis of the NER component of the TXM natural language processing (NLP) pipeline designed to detect entity relations and normalisations (Grover et al., 2007). The tagger, in common with many ML approaches to NER, reduces the entity recognition problem to a sequence tagging problem by using the BIO encoding of entities. As well as performing well on the CoNLL-2003 task, Maximum Entropy Markov Models have also been successful on biomedical NER tasks (Finkel et al., 2005). As the vanilla C&C tagger (Curran and Clark, 2003) is optimised for performance on newswire text, various modifications were applied to improve its performance for biomedical NER. Table 3 lists the extra features specifically designed for biomedical text. The C&C tagger was also extended using several gazetteers, including a protein, complex, experimental method and modification gazetteer, targeted at recognising entities occurring in the EPPI data. Further post-processing specific to the EPPI data involves correcting boundaries of some hyphenated proteins and filtering out entities ending in punctuation.

All experiments with the C&C tagger involve 5-fold cross-validation on all 2,000 GENIA abstracts and the combined EPPI TRAIN and DEVTEST sets. Cross-validation is carried out at the document level. For simple tagging, the C&C tagger is trained on the non-containing entities (innermost) or on the non-embedded entities (outermost). For inside-out and outside-in layering, a separate C&C model is trained for each layer of entities in the data, i.e. four models for the GENIA data and three models for the EPPI data. Cascading is performed on individual entities with different orderings, either ordering en-

Feature	Description
CHARACTER	Regular expressions matching typical protein names
WORDSHAPE	Extended version of the C&C WORDTYPE feature
HEADWORD	Head word of the current noun phrase
ABBREVIATION	Term identified as an abbreviation of a gazetteer term within a document
TITLE	Term seen in a noun phrase in the document title
WORDCOUNTER	Non-stop word that is among the 10 most frequent ones in a document
VERB	Verb lemma information added to each noun phrase token in the sentence
FONT	Text in italic and subscript contained in the original document format

Table 3: Extra features added to C&C .

tity models according to performance or entity frequency in the training data, ranging from highest to lowest. Cascading is also carried out on groups of entities (e.g. one model for all entities, one for a specific entity type, and combinations). Subsequent models in the cascade have access to the guesses of previous ones via a GUESS feature. Finally, joined label tagging is done by concatenating individual BIO tags from the innermost to the outermost layer.

As in the GENIA corpus, the most frequently annotated entity type in the EPPI data is protein with almost 55% of all annotations in the combined TRAIN and DEVTEST data (see Table 5). Given that the scores reported in this paper are calculated as $F1$ micro-averages over all categories, they are strongly influenced by the classifier's performance on proteins. However, scoring is not limited to a particular layer of entities (e.g. only outermost layer), but includes all levels of nesting. During scoring, a correct match is achieved when exactly the same sequence of text (encoded in start/end offsets) is marked with the same entity type in the gold standard and the system output. Precision, recall and $F1$ are calculated in standard fashion from the number of true positive, false positive and false negative NEs recognised.

6 Results

Table 4 lists overall cross-validation $F1$-scores calculated for all NEs at all levels of nesting when applying the various modelling techniques. For GENIA, cascading on individual entities when ordering entity models by performance yields the highest $F1$-score of 67.88. Using this method yields an increase of 3.26 $F1$ over the best simple tagging method, which scores 64.62 $F1$. Joined label tagging results in the second best overall $F1$-score of 67.82. Both layering (inside-out) and cascading (combining a model trained on all NEs with 4 models trained on other name, DNA, protein, or RNA) also perform competitively, reaching $F1$-scores of 67.62 and 67.56, respectively. In the experiments with the EPPI corpus, cascading is also the winner with an $F1$-score of 70.50 when combining a model trained on all NEswith one trained on fusions. This method only results in a small, yet statistically significant (χ^2, $p \leq 0.05$), increase in $F1$ of 0.43 over the best simple tagging algorithm. This could be due to the smaller number of nested NEs in the EPPI data and the fact that this data contains many NEs with more than one category. Layering (inside-out) performs almost as well as cascading ($F1$=70.44).

The difference in the overall performance between the GENIA and the EPPI corpus is partially due to the difference in the number of NEs which C&C is required to recognise, but also due to the fact that all features used are optimised for the EPPI data and simply applied to the GENIA corpus. The only feature not used for the experiments with the GENIA corpus is FONT, as this information is not preserved in the original XML of that corpus.

7 Discussion and Conclusion

According to the results for the modelling techniques, each proposed method outperforms simple tagging. Cascading yields the best result on the GENIA ($F1$=67.88) and EPPI data ($F1$=70.50), see Table 5 for individual entity scores. However, it involves extensive amounts of experimentation to determine the best model combination. The best setup for cascading is clearly data set dependent. With larger numbers of entity types annotated in a given corpus, it becomes increasingly impractical to exhaustively test all possible orders and combinations. Moreover, training and tagging times are lengthened as more models are combined in the cascade.

GENIA V3.02		EPPI	
Technique	$F1$	Technique	$F1$
Simple Tagging			
Training on innermost entities	64.62	Training on innermost entities	70.07
Training on outermost entities	62.72	Training on outermost entities	69.18
Layering			
Inside-out	*67.62*	Inside-out	*70.44*
Outside-in	*67.02*	Outside-in	70.21
Cascading			
Individual NE models (by performance)	***67.88***	Individual NE models (by performance)	70.42
Individual NE models (by frequency)	*67.72*	Individual NE models (by frequency)	*70.43*
All-cell_type	64.55	All-complex	70.03
All-DNA	*65.02*	All-drug/compound	70.08
All-other_name	*66.99*	All-fusion	***70.50***
All-protein	64.77	All-protein	70.02
All-RNA	64.80	All-complex-fusion	*70.46*
All-other_name-DNA-protein-RNA	*67.56*	All-drug/compound-fusion	*70.50*
Joined label tagging			
Inside-out	*67.82*	Inside-out	70.37

Table 4: Cross-validation $F1$-scores for different modelling techniques on the GENIA and EPPI data. Scores in italics mark statistically significant improvements (χ^2, $p \leq 0.05$) over the best simple tagging score.

Despite the large number of tags involved in using joined label tagging, this method outperforms simple tagging for both data sets and even results in the second-best overall $F1$-score of 67.72 obtained for the GENIA corpus. The fact that joined label tagging only requires training and tagging with one model makes this approach a viable alternative to cascading which is far more time-consuming to run.

Inside-out layering performs competitively both for the GENIA corpus ($F1$=67.62) and the EPPI corpus ($F1$=70.37), considering how little time is involved in setting up such experiments. As with joined label tagging, minimal optimisation is required when using this method. One disadvantage (as compared to simple, and to some extent joined label tagging) is that training and tagging times increase with the number of layers that are modelled.

In conclusion, this paper introduced and tested three different modelling techniques for recognising nested NEs, namely layering, cascading, and joined label tagging. As each of them reduces nested NER to one or more BIO-encoding problems, a conventional sequence tagger can be used. It was shown that each modelling technique outperfoms the simple tagging method for both biomedical data sets.

Future work will involve testing the proposed techniques on other data sets containing entity nesting, including the ACE data. We will also determine their merit when applying a different learning algorithm. Furthermore, possible solutions for recognising discontinuous entities will be investigated.

8 Acknowledgements

The authors are very grateful to the annotation team, and to Cognia (`http://www.cognia.com`) for their collaboration on the TXM project. This work is supported by the Text Mining Programme of ITI Life Sciences Scotland (`http://www.itilifesciences.com`).

References

K. Bretonnel Cohen, Lynne Fox, Philip V. Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics*, pages 38–45.

James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of CoNLL-2003*, pages 164–167.

GENIA V3.02					EPPI				
Entity type	Count	P	R	$F1$	Entity type	Count	P	R	$F1$
All	94,014	69.3	66.5	67.9	All	134,059	73.1	68.1	70.5
protein	34,813	75.1	74.9	75.0	protein	73,117	76.2	82.1	79.0
other name	20,914	60.0	67.2	63.4	expt. method	12,550	74.3	72.4	73.3
DNA	10,589	64.2	57.5	60.6	fragment	11,571	54.5	41.7	47.3
cell type	7,408	71.2	69.2	70.2	drug/compound	10,236	64.9	37.7	47.7
other org. compound	4,109	76.6	57.8	65.9	cell line	6,505	68.3	53.4	59.9
cell line	4,081	66.3	53.8	59.4	complex	6,454	62.5	32.2	42.5
lipid	2,359	76.9	65.6	70.8	modification	5,727	95.4	94.2	94.8
virus	2,133	76.0	73.4	74.7	mutant	4,025	40.7	23.2	29.6
multi-cell	1,784	72.5	60.1	65.7	fusion	3,874	56.6	36.0	44.0

Table 5: Individual counts and scores of the most frequent GENIA and all EPPI entity types for the best-performing method: cascading.

Jenny Rose Finkel, Shipra Dingare, Christopher D. Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. 2005. Exploring the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics*, 6(Suppl1):S5.

Claire Grover and Richard Tobin. 2006. Rule-based chunking and reusability. In *Proceedings of LREC 2006*, pages 873–878.

Claire Grover, Michael Matthews, and Richard Tobin. 2006. Tools to address the interdependence between tokenisation and standoff annotation. In *Proceedings of NLPXML 2006*, pages 19–26.

Claire Grover, Barry Haddow, Ewan Klein, Michael Matthews, Leif Arda Nielsen, Richard Tobin, and Xinglong Wang. 2007. Adapting a relation extraction pipeline for the BioCreAtIvE II task. In *Proceedings of the BioCreAtIvE Workshop 2007*, Madrid, Spain.

Baohua Gu. 2006. Recognizing nested named entities in GENIA corpus. In *Proceedings of the BioNLP Worshop, HLT-NAACL 2006*, pages 112–113.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of JNLPBA 2004*, pages 70–75.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, and Lyle Ungar. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of BioLINK 2004*, pages 61–68.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *Proceedings of HLT/EMNLP 2005*, pages 987–994.

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of INLG 2000*, pages 201–208.

Tomoko Ohta, Yuka Tateisi, Hideki Mima, and Jun'ichi Tsujii. 2002. GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of HLT 2002*, pages 73–77.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora (ACL 1995)*, pages 82–94.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, pages 451–462.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the BioNLP Workshop, ACL 2003*, pages 49–56.

Larry Smith, Tom Rindflesch, and W. John Wilbur. 2004. Med-Post: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147.

Jie Zhang, Dan Shen, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2004. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37(6):411–422.

Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.

Guodong Zhou. 2006. Recognizing names in biomedical texts using mutual information independence model and svm plus sigmoid. *International Journal of Medical Informatics*, 75:456–467.