Enhancing Medical Named Entity Recognition with Features Derived from Unsupervised Methods

Maria Skeppstedt
Dept. of Computer and Systems Sciences (DSV)
Stockholm University, Forum 100, 164 40 Kista, Sweden
mariask@dsv.su.se

Abstract

A study of the usefulness of features extracted from unsupervised methods is proposed. The usefulness of these features will be studied on the task of performing named entity recognition within one clinical sub-domain as well as on the task of adapting a named entity recognition model to a new clinical sub-domain. Four named entity types, all very relevant for clinical information extraction, will be studied: Disorder, Finding, Pharmaceutical Drug and Body Structure. The named entity recognition will be performed using conditional random fields. As unsupervised features, a clustering of the semantic representation of words obtained from a random indexing word space will be used.

1 Introduction

Creating the annotated corpus needed for training a NER (named entity recognition) model is costly. This is particularly the case for texts in specialised domains, for which expert annotators are often required. In addition, the need for expert annotators also limits the possibilities of using crowdsourcing approaches (e.g. Amazon Mechanical Turk). Features from unsupervised machine-learning methods, for which no labelled training data is required, have, however, been shown to improve the performance of NER systems (Jonnalagadda et al., 2012). It is therefore likely that by incorporating features from unsupervised methods, it is possible to reduce the amount of training data needed to achieve a fixed level of performance.

Due to differences in the use of language, an NLP system developed for, or trained on, text from one sub-domain often shows a drop in performance when applied on texts from another subdomain (Martinez et al., 2013). This has the effect that when performing NER on a new subdomain, annotated text from this new targeted subdomain might be required, even when there are annotated corpora from other domains. It would, however, be preferable to be able to apply a NER model trained on text from one sub-domain on another sub-domain, with only a minimum of additional data from this other targeted sub-domain. Incorporating features from unsupervised methods might limit the amount of additional annotated data needed for adapting a NER model to a new sub-domain.

The proposed study aims at investigating the usefulness of unsupervised features, both for NER within one sub-domain and for domain adaptation of a NER model. The study has two hypotheses.

- Within one subdomain:

  For reaching the same level of performance when training a NER model, less training data is required when unsupervised features are used.

- For adapting a model trained on one subdomain to a new targeted subdomain:

  For reaching the same level of performance when adapting a NER model to a new subdomain, less additional training data is required in the new targeted subdomain when unsupervised features are used.

For both hypotheses, the level of performance is defined in terms of F-score.

The proposed study will be carried out on different sub-domains within the specialised text domain of *clinical text*.

2 Related research

There are a number of previous studies on named entity recognition in clinical text. For instance, a corpus annotated for the entities Condition,

Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 21–30,
Gothenburg, Sweden, April 26-30 2014.

Drug/Device and Locus was used for training a support vector machine with uneven margins (Roberts et al., 2008) and a corpus annotated for the entities Finding, Substance and Body was used for training a conditional random fields (CRF) system (Wang, 2009) as well as for training an ensemble of different classifiers (Wang and Patrick, 2009). Most studies have, however, been conducted on the *i2b2 medication challenge* corpus and the *i2b2 challenge on concepts, assertions, and relations* corpus. Conditional random fields (Patrick and Li, 2010) as well as an ensemble classifier (Doan et al., 2012) has for instance been used for extracting the entity Medication names from the *medication challenge* corpus, while all but the best among the top-performing systems used CRF for extracting the entities Medical Problem, Test and Treatment from the *i2b2 challenge on concepts, assertions, and relations* corpus (Uzuner et al., 2011). The best system (de Bruijn et al., 2011) used semi-Markov HMM, and in addition to the features used by most of the other systems (e.g. tokens/lemmas/stems, orthographics, affixes, part-of-speech, output of terminology matching), this system also used features extracted from hierarchical word clusters on un-annotated text. For constructing the clusters, they used Brown clustering, and represented the feature as a 7-bit showing to what cluster a word belonged.

Outside of the biomedical domain, there are many studies on English corpora, which have shown that using features extracted from clusters constructed on unlabelled corpora improves performance of NER models, especially when using a smaller amount of training data (Miller et al., 2004; Freitag, 2004). This approach has also been shown to be successful for named entity recognition in other languages, e.g. German, Dutch and Spanish (Täckström et al., 2012), as well as on related NLP tasks (Biemann et al., 2007), and there are NER tools that automatically incorporate features extracted from unsupervised methods (Stanford, 2012). There are a number of additional studies within the biomedical domain, e.g. using features from Brown and other clustering approaches (Stenetorp et al., 2012) or from k-means clustered vectors from a neural networks-based word space implementation (Pyysalo et al., 2014). Jonnalagadda et al. (2012) also present a study in which unsupervised features are used for training a model on the *i2b2 challenge on concepts, assertions, and relations* corpus. As unannotated corpus, they used a corpus created by extracting Medline abstracts that are indexed with the publication type "clinical trials". They then built a semantic representation of this corpus in the form of a random indexing-based word space. This representation was then used for extracting a number of similar words to each word in the *i2b2 challenge on concepts, assertions, and relations* corpus, which were used as features when training a CRF system. The parameters of the random indexing model were selected by letting the nearest neighbours of a word vote for one of the UMLS categories Medical Problem, Treatment and Test according to the category of the neighbour, and by comparing the category winning the vote to the actual category of the word. The authors motivate their choice of using random indexing for creating features with that this method is scalable to very large corpora without requiring large computational resources.

The method proposed here is similar to the method used by Jonnalagadda et al. (2012). However, the focus of the proposed study is to explore to what extent unsupervised features can help a machine learning system trained only on very little data. It is therefore not feasible to use the large number of features that would be generated by using neighbouring words, as that would require a large training data set to ensure that there are enough training examples for each generated feature. Therefore, the proposed method instead further processes the word space model by constructing clusters of semantically related words, thereby reducing the number of generated features, similar to the approach by Pyysalo et al. (2014).

3 Materials and previous results

Texts from three different clinical sub-domains: *cardiac ICU* (intensive care unit), *orthopaedic ER* (emergency room), and *internal medicine ER* have been annotated (Tables 1-3).[1] All texts are written in Swedish, and they all share the characteristics of text types written under time pressure; all of them containing many abbreviations and incomplete sentences. There are, however, also differences in e.g. what abbreviations are used and what

[1]Research on these texts aiming at extracting information related to Disorders/Findings and Pharmaceutical Drugs has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5.

Data set:	All	
Entity category	# entities	(Unique)
Disorder	1088	(533)
Finding	1798	(1295)
Pharmaceuticals	1048	(497)
Body structure	461	(252)

Table 1: Annotated data, Cardiac ICU

Data set:	All	
Entity category	# entities	(Unique)
Disorder	1258	(541)
Finding	1439	(785)
Pharmaceuticals	880	(212)
Body structure	1324	(423)

Table 2: Annotated data, Orthopaedic ER

entities that are frequently mentioned.

The texts from cardiac ICU and orthopaedic ER will be treated as existing annotations in a current domain, whereas internal medicine ER will be treated as the new target domain. Approximately a third of the texts from internal medicine ER have been doubly annotated, and an evaluation set has been created by manually resolving differences between the two annotators (Skeppstedt et al., 2014). This evaluation subset will be used as held-out data for evaluating the NER task.

The following four entity categories have been annotated (Skeppstedt et al., 2014): (1) Disorder (a disease or abnormal condition that is not momentary and that has an underlying pathological process), (2) Finding (a symptom reported by the patient, an observation made by the physician or the result of a medical examination of the patient), (3) Pharmaceutical Drug (not limited to generic name or trade name, but includes also e.g. drugs expressed by their effect, such as painkiller or sleeping pill). (4) Body Structure (an anatomically defined body part).

These three annotated corpora will be used in the proposed study, together with a large corpus of un-annotated text from which unsupervised features will be extracted. This large corpus will be a subset of the Stockholm EPR corpus (Dalianis et al., 2009), which is a large corpus of clinical text written in Swedish.

Named entity recognition on the internal medicine ER part of the annotated corpus has already been studied, and results on the evaluation set were an F-score of 0.81 for the entity Disorder, 0.69 for Finding, 0.88 for Pharmaceutical Drug, 0.85 for Body Structure and 0.78 for the combined category Disorder + Finding (Skeppstedt et al., 2014). Features used for training the model on the development/training part of the internal medicine ER corpus were the lemma forms of the words, their part of speech, their semantic category in used vocabulary lists, their word constituents (if the words were compounds) as well as the orthographics of the words. A narrow context window was used, as shown by the entries marked in boldface in Figure 1. As terminologies, the Swedish versions of SNOMED CT[2], MeSH[3], ICD-10[4], the Swedish medical list FASS[5] were used, as well as a vocabulary list of non-medical words, compiled from the Swedish Parole corpus (Gellerstam et al., 2000).

4 Methodological background

The proposed method consists of using the training data first for parameter setting (through n-fold cross-validation) and thereafter for training a model using the best parameters. This model is then to be evaluated on held-out data. A number of rounds with parameter setting and training will be carried out, where each new round will make use of an increasingly larger subset of the training data. Two versions of parameter setting and model training will be carried out for each round; one using features obtained from unsupervised methods on un-annotated text and one in which such features are not used. The results of the two versions are then to be compared, with the hypothesis that the model incorporating unsupervised methods will perform better, at least for small training data sizes.

To accomplish this, the proposed method makes use of four main components: (1) A system for training a NER model given features extracted from an annotated corpus. As this component, a conditional random fields (CRF) system will be used. (2) A system for automatic parameter setting. As a large number of models are to be constructed on different sizes of the training data, for which optimal parameters are likely to differ, parameters for each set of training data has to be determined automatically for it to be feasible to

[2] www.ihtsdo.org
[3] mesh.kib.ki.se
[4] www.who.int/classifications/icd/en/
[5] www.fass.se

Data set:	Development		Final evaluation
Entity category	# entities	(Unique)	# entities
Disorder	1,317	(607)	681
Finding	2,540	(1,353)	1282
Pharmaceuticals	959	(350)	580
Body structure	497	(197)	253
Tokens in corpus	45,482		25,370

Table 3: Annotated entities, internal medicine ER

Token	*Lemma*	*POS*	*Terminology*	*Compound*		*Orthographics*	*Cluster membership level 1*	..	*Cluster membership level n*	*Category*
DVT	*dvt*	*noun*	*disorder*	-	-	*all upper*	*#40*	..	*#39423*	*B-Disorder*
patient	*patient*	**noun**	*person*	-	-	-	*#3*	..	*#23498*	*O*
with	**with**	**prep.**	**parole**	-	-	-	*#14*	..	*#30892*	*O*
chestpain	**chestpain**	**noun**	**finding**	**chest**	**pain**	-	*#40*	..	*#23409*	*B-Finding* ← **Current**
and	*and*	**conj.**	*parole*	-	-	-	-	..	-	*O*
problems	*problem*	*noun*	*finding*	-	-	-	*#40*	..	*#23409*	*B-Finding*
to	*to*	*prep.*	*finding*	-	-	-	-	..	-	*I-Finding*
breathe	*breathe*	*verb*	*finding*	-	-	-	*#90*	..	*#23409*	*I-Finding*

Figure 1: A hypothetical example sentence, with hypothetical features for training a machine learning model. Features used in a previous medical named entity recognition study (Skeppstedt et al., 2014) on this corpus are shown in boldface. The last column contains the entity category according to the manual annotation.

carry out the experiments. (3) A system for representing semantic similarity of the words in the un-annotated corpus. As this component, a random indexing based word space model will used. (4) A system for turning the semantic representation of the word space model into features to use for the NER model. As this component, clustering will be used.

To give a methodological background, the theoretical foundation for the four components will be described.

4.1 Conditional random fields

Conditional random fields (CRF or CRFs), introduced by Lafferty et al. (2001), is a machine learning method suitable for segmenting and labelling sequential data and therefore often used for e.g. named entity recognition. As described in the related research section, CRFs have been used in a number of studies for extracting entities from clinical text. In contrast to many other types of data, observed data points for sequential data, such as text, are dependent on other observed data points. Such dependences between data points are practical to describe within the framework of graphical models (Bishop, 2006, p. 359), to which CRF belongs (Sutton and McCallum, 2006, p. 1). In the special, but frequently used, case of linear chain CRF, the output variables are linked in a chain. Apart from being dependent on the input variables, each output variable is then conditionally independent on all other output variables, except on the previous and following output variable, given these two neighbouring output variables. In a named entity recognition task, the output variables are the named entity classes that are to be predicted and the observed input variables are observed features of the text, such as the tokens or their part-of-speech.

CRF is closely related to Hidden Markov Models, which is also typically described as a graphical model. A difference, however, is that Hidden Markov Models belongs to the class of generative models, whereas CRF is a conditional model (Sutton and McCallum, 2006, p. 1). Generative models model the joint distribution between input variables and the variables that are to be predicted (Bishop, 2006, p. 43). In contrast, CRF and other conditional models instead directly model the conditional distribution, enabling the use of a larger

feature set (Sutton and McCallum, 2006, p. 1).

For named entity recognition, the IOB-encoding is typically used for encoding the output variables. Tokens not annotated as an entity are then encoded with the label *O*, whereas labels for annotated tokens are prefixed with a *B*, if it is the first token in the annotated chunk, and an *I* otherwise (Jurafsky and Martin, 2008, pp. 763–764). An example of this encoding is shown in the last column in Figure 1. In this case, where there are four types of entities, the model thus learns to classify in 8+1 different classes: B-Disorder, I-Disorder, B-Finding, I-Finding, B-Drug, I-Drug, B-BodyStructure, I-BodyStructure and *O*.

The dependencies are defined by a large number of (typically binary) feature functions of input and output variables. E.g. is all of the following true?

- Output: The output at the current position is **I-Disorder**
- Output: The output at the previous position is **B-Disorder**
- Input: The token at the current position is **chest-pain**
- Input: The token at the previous position is **experiences**

A feature function in a linear chain CRF can only include the values of the output variable in current position and in the immediate previous position, whereas it can include, and thereby show a dependence on, input variables from any position.

The CRF model is trained through setting weights for the feature functions, which is carried out by penalised maximum likelihood. *Penalised* means that regularisation is used, and regularisation is performed by adding a penalty term, which prevents the weights from reaching too large values, and thereby prevents over-fitting (Bishop, 2006, p. 10). The L1-norm and the L2-norm are frequently used for regularisation (Tsuruoka et al., 2009), and a variable C governs the importance of the regularisation. Using the L1-norm also results in that if C is large enough, some of the weights are driven to zero, resulting in a sparse model and thereby the feature functions that those weights control will not play any role in the model. Thereby, complex models can be trained also on data sets with a limited size, without being over-fitted. However, a suitable value of C must still be determined (Bishop, 2006, p. 145).

The plan for the proposed study is to use the CRF package CRF++[6], which has been used in a number of previous NER studies, also in the medical domain. The CRF++ package automatically generates feature functions from user-defined templates. When using CRF++ as a linear chain CRF, it generates one binary feature function for each combination of output class, previous output class and unique string in the training data that is expanded by a template. This means that L * L * M feature functions are generated for each template, where L = the number of output classes and M = the number of unique expanded strings. If only the current token were to be used as a feature, the number of feature functions would be $9*9*|unique\ tokens\ in\ the\ corpus|$. In practice, a lot of other features are, however, used. Most of these features will be of no use to the classifier, which means that it is important to use an inference method that sets the weights of the feature functions with irrelevant features to zero, thus an inference method that promotes sparsity.

4.2 Parameter setting

As previously explained, a large number of models are to be constructed, which requires a simple and efficient method for parameter setting. An advantage with using the L1-norm is that only one parameter, the C-value, has to be optimised, as the weights for feature functions are driven to zero for feature functions that are not useful. The L1-norm will therefore be used in the proposed study. A very large feature set can then be used, without running the risk of over-fitting the model. Features will include those that have been used in previous clinical NER studies (Jonnalagadda et al., 2012; de Bruijn et al., 2011; Skeppstedt et al., 2014), with a context window of four previous and four following tokens.

When maximising the conditional log likelihood of the parameters, the CRF++ program will set parameters that are optimal for training the model for the best micro-averaged results for the four classes Disorder, Finding, Pharmaceutical drug and Body structure. A hill climbing search (Marsland, 2009, pp. 262–264) for finding a good C-value will be used, starting with a value very close to zero and thereafter changing it in a direction that improves the NER results. A decreasingly smaller step size will be used for changing

[6] crfpp.sourceforge.net

Lemmatised and stop word filtered with a window size of 2 (1+1):

	complain	dermatitis	eczema	itch	patient
complain:	[0	0	0	2	2]
dermatitis:	[0	0	0	1	0]
eczema:	[0	0	0	1	0]
itch:	[2	1	1	0	0]
patient:	[2	0	0	0	0]

Figure 2: Term-by-term co-occurrence matrix for the small corpus "Patient complains of itching dermatitis. Patient complains of itching eczema."

the C-value, until only small changes in the results can be observed.

4.3 Random indexing

Random indexing is one version of the word space model, and as all word space models it is a method for representing distributional semantics. The random indexing method was originally devised by Kanerva et al. (2000), to deal with the performance problems (in terms of memory and computation time) that were associated with the LSA/LSI implementations at that time. Due to its computational efficiency, random indexing remains to be a popular method when building distributional semantics models on very large corpora, e.g. large web corpora (Sahlgren and Karlgren, 2009) or Medline abstracts (Jonnalagadda et al., 2012).

Distributional semantics is built on the distributional hypothesis, which states that "Words with similar meanings tend to occur in similar contexts". If *dermatitis* and *eczema* often occur in similar contexts, e.g. "Patient complains of itching *dermatitis*" and "Patient complains of itching *eczema*", it is likely that *dermatitis* and *eczema* have a similar meaning. One possible method of representing word co-occurrence information is to construct a term-by-term co-occurrence matrix, i.e. a matrix of dimensionality $w \times w$, in which w is the number of terms (unique semantic units, e.g. words) in the corpus. The elements of the matrix then contain the number of times each semantic unit occurs in the context of each other semantic unit (figure 2).

The *context vectors* of two semantic units can then be compared as a measure of semantic similarity between units, e.g. using the the euclidian distance between normalised context vectors or the cosine similarity.

	1	**2**	**3**	**...**	**d**
...	[0	0	1	...	0]
complain:	[0	0	0	...	1]
itch:	[0	1	1	...	0]
patient:	[-1	0	0	...	0]
...	[...	...	...	...	..]
word w	[0	0	-1	...	0]

Figure 3: Index vectors.

The large dimension of a term-by-term matrix leads, however, to scalability problems, and the typical solution to this is to apply dimensionality reduction on the matrix. In a semantic space created by latent semantic analysis, for instance, dimensionality reduction is performed by applying the linear algebra matrix operation singular value decomposition (Landauer and Dutnais, 1997). *Random indexing* is another solution, in which a matrix with a smaller dimension is created from start, using the following method (Sahlgren et al., 2008):

Each term in the data is assigned a unique representation, called an *index vector*. The index vectors all have the dimensionality d (where $d \geq 1000$ but $\ll w$). Most of the elements of the index vectors are set to 0, but a few, randomly selected, elements are set to either +1 or -1. (Usually around 1-2% of the elements.) Instead of having orthogonal vectors, as is the case for the term-by-term matrix, the index vectors are nearly orthogonal. (See Figure 3.)

Each term in the data is also assigned a *context vector*, also of the dimensionality d. Initially, all elements in the context vectors are set to 0. The context vector of each term is then updated by, for every occurrence of the term in the corpus, adding the index vectors of the neighboring words. The neighboring words are called the *context window*, and this can be both narrow or wide, depending on what semantic relations the word space model is intended to capture. The size of the context window can have large impact on the results (Sahlgren et al., 2008), and for detecting paradigmatic relations (i.e. words that occur in similar contexts, rather than words that occur together) a fairly narrow context window has been shown to be most effective.

The resulting context vectors form a matrix of dimension $w \times d$. This matrix is an approximation of the term-by-term matrix, and the same similar-

Index vectors (never change)

	1	2	3	...	d
...					
itching:	[0	1	1	...	0]
patient:	[-1	0	0	...	0]
...					

Context vectors

	1	2	3	...	d
...					
complain:	[-1	1	1	...	0]
...					

Figure 4: The updated context vectors.

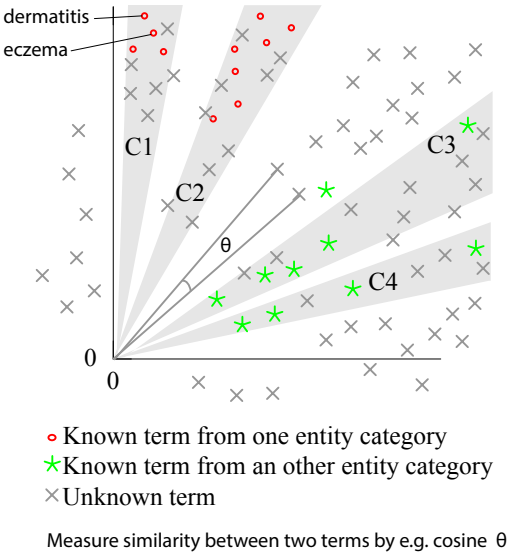

Figure 5: Context vectors for terms in a hypothetical word space with d=2. The context vectors for the semantically similar words *eczema* and *dermatitis* are close in the word space, in which closeness is measured as the cosine of the angle between the vectors. Four hypothetical clusters (C1-C4) of context vectors are also shown; clusters that contain a large proportion of known terms.

ity measures can be applied.

A hypothetical word space with d=2 is shown in Figure 5.

4.4 Clustering

As mentioned earlier, for the word space information to be useful for training a CRF model on a small data set, it must be represented as a feature that can only take a limited number of different values. The proposed methods for achieving this is to cluster the context vectors of the word space model, similar to what has been done in previous research (Pyysalo et al., 2014). Also similar to previous research, cluster membership for a word in the NER training and test data will be used as a feature. Four named hypothetical clusters of context vectors are shown in the word space model in Figure 5 to illustrate the general idea, and an example of how to use cluster membership as a feature is shown Figure 1.

Different clustering techniques will be evaluated, for the quality of the created clusters, as well as for their computational efficiency. Having hierarchical clusters might be preferable, as cluster membership to clusters of different granularity then can be offered as features for training the CRF model. Which granularity that is most suitable might vary depending on the entity type and also depending on the size of the training data. However, e.g. performing hierarchical agglomerative clustering (Jurafsky and Martin, 2008, p. 700) on the entire unlabelled corpus might be computationally intractable (thereby defeating the purpose of using random indexing), as it requires pairwise comparisons between the words in the corpus. The pairwise comparison is a part of the agglomerative clustering algorithm, in which each word is first assigned its own cluster and then each pair of clusters is compared for similarity, resulting in a merge of the most similar clusters. This process is thereafter iteratively repeated, having the distance between the centroids of the clusters as similarity measure. An alternative, which requires a less efficient clustering algorithm, would be to not create clusters of all the words in the corpus, but to limit initially created clusters to include those words that occur in available terminologies. Cluster membership of unknown words in the corpus could then be determined by measuring similarity to the centroids of these initially created clusters.

Regardless of what clustering technique that is chosen, the parameters of the random indexing models, as well as of the clustering, will be determined by evaluating to what extent words that belong to one of the studied semantic categories (according to available terminologies) are clustered together. This will be measured using *purity* and *inverse purity* (Amigó et al., 2009). However, if clusters are to be created from all words in the corpus, the true semantic category will only be known for a very small subset of clustered words. In that case, the two measures have to be defined as *purity* being to what extent a cluster only contains known words of one category and *inverse purity* being the extent to which known words of the same category are grouped into the same cluster.

5 Proposed experiments

The first phase of the experiments will consist of finding the best parameters for the random indexing model and the clustering, as described above.

The second phase will consist of evaluating the usefulness of the clustered data for the NER task. Three main experiments will be carried out in this phase (I, II and III), using data set(s) from the following sources:

I: Internal medicine ER

II: Internal medicine ER + Cardiac ICU

III: Internal medicine ER + Orthopaedic ER

In each experiment, the following will be carried out:

1. Divide internal medicine ER training data into 5 partitions (into a random division, to better simulate the situation when not all data is available, using the same random division for all experiments).
2. Run step 3-5 in 5 rounds. Each new round uses one additional internal medicine ER partition: (Experiments II and III always use the entire data set from the other domain). In each round, two versions of step 3-5 will be carried out:
   (a) With unsupervised features.
   (b) Without unsupervised features.
3. Use training data for determining C-value (by n-fold cross-validation).
4. Use training data for training a model with this C-value.
5. Evaluate the model on the held-out internal medicine ICU data.

6 Open issues

What clustering technique to use has previously been mentioned as one important open issue. The following are examples of other open issues:

- Could the information obtained from random indexing be used in some other way than as transformed to cluster membership features? Jonnalagadda et al. (2012) used the terms closest in the semantic space as a feature. Could this method be adapted in some way to models constructed with a small amount of training data? For instance by restricting what terms are allowed to be used as such a feature, and thereby limiting the number of possible values this feature can take.
- Would it be better to use other approaches (or compare different approaches) for obtaining features from unlabelled data? A possibility could be to use a more standard clustering approach, such as Brown clustering used in previous clinical NER studies (de Bruijn et al., 2011). Another possibility could be to keep the idea of creating clusters from vectors in a word space model, but to use other methods than random indexing for constructing the word space; e.g. the previously mentioned latent semantic analysis (Landauer and Dutnais, 1997), or a neural networks-based word space implementation (Pyysalo et al., 2014).
- Many relevant terms within the medical domain are multi-word terms (e.g. of the type *diabetes mellitus*), and there are studies on how to construct semantic spaces with such multiword terms as the smallest semantic unit (Henriksson et al., 2013). Should the whitespace segmented token be treated as the smallest semantic unit in the proposed study, or should the use of larger semantic units be considered?

Acknowledgements

Many thanks to Aron Henriksson, Alyaa Alfalahi, Maria Kvist, Gunnar Nilsson and Hercules Dalianis for taking a very active part in the planing of the proposed study, as well as to the three anonymous reviewers for their constructive and detailed comments.

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486, aug.

Chris Biemann, Claudio Giuliano, and Alfio Gliozzo. 2007. Unsupervised part of speech tagging supporting supervised methods. In *RANLP*.

Christopher M. Bishop. 2006. *Pattern recognition and machine learning*. Springer, New York, NY.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research, Kalmar, Sweden*, pages 243–249.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel D. Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc*, 18(5):557–562.

Son Doan, Nigel Collier, Hua Xu, Hoang Duy Pham, and Minh Phuong Tu. 2012. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC Med Inform Decis Mak*, 12:36.

Dayne Freitag. 2004. Trained named entity recognition using distributional clusters. In *EMNLP*, pages 262–269.

M Gellerstam, Y Cederholm, and T Rasmark. 2000. The bank of Swedish. In *LREC 2000. The 2nd International Conference on Language Resources and Evaluation*, pages 329–333, Athens, Greece.

Aron Henriksson, Mike Conway, Martin Duneld, and Wendy W. Chapman. 2013. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA 2013)*, Washington DC, USA.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. Enhancing clinical concept extraction with distributional semantics. *J Biomed Inform*, 45(1):129–40, Feb.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition, February.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In L. R. Gleitman and A. K. Joshi, editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah, NJ.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Thomas K Landauer and Susan T. Dutnais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Stephen Marsland. 2009. *Machine learning : an algorithmic perspective*. Chapman & Hall/CRC, Boca Raton, FL.

David Martinez, Lawrence Cavedon, and Graham Pitson. 2013. Stability of text mining techniques for identifying cancer staging. In *Proceedings of the 4th International Louhi Workshop on Health Document Text Mining and Information Analysis - Louhi 2013*, Sydney, Australia, February.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT*, pages 337–342.

Jon Patrick and Min Li. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc*, 17(5):524–527, Sep-Oct.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2014. Distributional semantics resources for biomedical text processing. In *Proceedings of Languages in Biology and Medicine*.

Angus Roberts, Robert Gaizasukas, Mark Hepple, and Yikun Guo. 2008. Combining terminology resources and statistical methods for entity recognition: an evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2974–2979, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Magnus Sahlgren and Jussi Karlgren. 2009. Terminology mining in social media. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 1300–1305.

Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *J Biomed Inform*, Feb (in press).

NLP Group Stanford. 2012. Stanford Named Entity Recognizer (NER). http://www-nlp.stanford.edu/software/CRF-NER.shtml. Accessed 2012-03-29.

Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, and Takashi Chikayama. 2012. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine*.

Charles. Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada, June. Association for Computational Linguistics.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Fast full parsing by linear-chain conditional random fields. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 790–798, Stroudsburg, PA, USA. Association for Computational Linguistics.

Özlem. Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556.

Yefeng Wang and Jon Patrick. 2009. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 42–49.

Yefeng Wang. 2009. Annotating and recognising named entities in clinical notes. In *Proceedings of the ACL-IJCNLP Student Research Workshop*, pages 18–26, Singapore.