EACL - Expansion of Abbreviations in CLinical text

Lisa Tengstrand*, Beáta Megyesi*, Aron Henriksson⁺, Martin Duneld⁺ and Maria Kvist⁺
*Department of Linguistics and Philology,
Uppsala University, Sweden
tengstrand@ling.su.se, beata.megyesi@lingfil.uu.se
⁺Department of Computer and System Sciences,
Stockholm University, Sweden
aronhen@dsv.su.se, xmartin@dsv.su.se, maria.kvist@karolinska.se

Abstract

In the medical domain, especially in clinical texts, non-standard abbreviations are prevalent, which impairs readability for patients. To ease the understanding of the physicians' notes, abbreviations need to be identified and expanded to their original forms. We present a distributional semantic approach to find candidates of the original form of the abbreviation, and combine this with Levenshtein distance to choose the correct candidate among the semantically related words. We apply the method to radiology reports and medical journal texts, and compare the results to general Swedish. The results show that the correct expansion of the abbreviation can be found in 40% of the cases, an improvement by 24 percentage points compared to the baseline (0.16), and an increase by 22 percentage points compared to using word space models alone (0.18).

1 Introduction

Abbreviations are prevalent in text, especially in certain text types where the author has either limited space or time to write the written message and therefore shortens some words or phrases. This might, however, make it difficult for the reader to understand the meaning of the actual abbreviation. Although some abbreviations are well-known, and frequently used by most of us (e.g., i.e., pm, etc.), most of the abbreviations used in specialized domains are often less known to the public. Interpreting them is not an easy task, as abbreviations are often ambiguous and their correct meaning depends on the context in which they appear. For example, military and governmental staff would naturally read EACL as Emergency Action Checklist, people in the food and beverage business might think of the company name EACL, linguists would probably interpret it as the European Chapter of Chinese Linguistics, while computational linguists would generally claim that EACL stands for the European Chapter of the Association for Computational Linguistics. However, the readers of this particular article know, as the title suggests, that the intended meaning here is the *Expansion of Abbreviations in CLinical text*.

It has been shown that abbreviations are frequently occurring in various domains and genres, such as in historical documents, messages in social media, as well as in different registers used by specialists within a particular field of expertise. Clinical texts produced by health care personnel is an example of the latter. The clinical texts are communication artifacts, and the clinical setting requires that information is expressed in an efficient way, resulting in short telegraphic messages. Physicians and nurses need to document their work to describe findings, treatments and procedures precisely and compactly, often under time pressure.

In recent years, governments and health care actors have started making electronic health records accessible, not only to other caretakers, but also to patients in order to enable them to participate actively in their own health care processes. However, several studies have shown that patients have difficulties to comprehend their own health care reports and other medical texts due to the different linguistic features that characterize these, aswell as to medical jargon and technical terminology (Elhadad, 2006; Rudd et al., 1999; Keselman et al., 2007). It has also been shown that physicians rarely adapt their writing style in order to produce documents that are accessible to lay readers (Allvin, 2010). Besides the use of different terminologies and technical terms, an important obstacle for patients to comprehend medical texts is the frequent use of – for the patients unknown – ab-

Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014, pages 94–103,
Gothenburg, Sweden, April 26-30 2014.

breviations (Keselman et al., 2007; Adnan et al., 2010).

In health records, abbreviations, which constitute linguistic units that are inherently difficult to decode, are commonly used and often non standard (Skeppstedt, 2012). An important step in order to increase readability for lay readers is to translate abbreviated words into their corresponding full length words.

The aim of this study is to explore a distributional semantic approach combined with word normalization, measured by Levenshtein distance, to abbreviation expansion. Using distributional semantic models, which can be applied to large amounts of data, has been shown to be a viable approach to extracting candidates for the underlying, original word of an abbreviation. In order to find the correct expansion among the semantically related candidates, we apply the Levenshtein distance measure. We report on experiments on comparative studies of various text types in Swedish, including radiology reports, medical journals and texts taken from a corpus of general Swedish.

2 Background

An abbreviation is a shorter – abbreviated – form of a word or phrase, often originating from a technical term or a named entity. Abbreviations are typically formed in one of three ways: by (i) clipping the last character sequence of the word (e.g., *pat* for *patient* or *pathology*), (ii) merging the initial letter(s) of the words to form an acronym (e.g., *UU* for *Uppsala University*), or (iii) merging some of the letters – often the initial letter of the syllables – in the word (e.g., *msg* for *message*). Abbreviations can also be formed as a combination of these three categories (e.g., *EACL* for *Expansion of Abbreviations in CLinical text*).

Automatically expanding abbreviations to their original form has been of interest to computational linguists as a means to improve text-to-speech, information retrieval and information extraction systems. Rule-based systems as well as statistical and machine learning methods have been proposed to detect and expand abbreviations. A common component of most solutions is their reliance on the assumption that an abbreviation and its corresponding definition will appear in the same text.

Taghva and Gilbreth (1999) present a method for automatic acronym-definition extraction in technical literature, where acronym detection is based on case and token length constraints. The surrounding text is subsequently searched for possible definitions corresponding to the detected acronym using an inexact pattern-matching algorithm. The resulting set of candidate definitions is then narrowed down by applying the Longest Common Subsequence (LCS) algorithm (Nakatsu et al., 1982) to the candidate pairs. They report 98% precision and 93% recall when excluding acronyms of two or fewer characters.

Park and Byrd (2001), along somewhat similar lines, propose a hybrid text mining approach for abbreviation expansion in technical literature. Orthographic constraints and stop lists are first used to detect abbreviations; candidate definitions are then extracted from the adjacent text based on a set of pre-specified conditions. The abbreviations and definitions are converted into patterns, for which transformation rules are constructed. An initial rule-base comprising the most frequent rules is subsequently employed for automatic abbreviation expansion. They report 98% precision and 94% recall as an average over three document types.

In the medical domain, most approaches to abbreviation resolution also rely on the co-occurrence of abbreviations and definitions in a text, typically by exploiting the fact that abbreviations are sometimes defined on their first mention. These studies extract candidate abbreviation-definition pairs by assuming that either the definition or the abbreviation is written in parentheses (Schwartz and Hearst, 2003). The process of determining which of the extracted abbreviation-definition pairs are likely to be correct is then performed either by rule-based (Ao and Takagi, 2005) or machine learning (Chang et al., 2002; Movshovitz-Attias and Cohen, 2012) methods. Most of these studies have been conducted on English corpora; however, there is one study on Swedish medical text (Dannélls, 2006). There are problems with this popular approach to abbreviation expansion: Yu et al. (2002) found that around 75% of all abbreviations in the biomedical literature are never defined.

The application of this method to clinical text is even more problematic, as it seems highly unlikely that abbreviations would be defined in this way. The telegraphic style of clinical narrative, with its many non-standard abbreviations, is reasonably explained by time constraints in the clinical setting. There has been some work on iden-

tifying such undefined abbreviations in clinical text (Isenius et al., 2012), as well as on finding the intended abbreviation expansion among candidates in an abbreviation dictionary (Gaudan et al., 2005).

Henriksson et al. (2012; 2014) present a method for expanding abbreviations in clinical text that does not require abbreviations to be defined, or even co-occur, in the text. The method is based on distributional semantic models by effectively treating abbreviations and their corresponding definition as synonymous, at least in the sense of sharing distributional properties. Distributional semantics (see Cohen and Widdows (2009) for an overview) is based on the observation that words that occur in similar contexts tend to be semantically related (Harris, 1954). These relationships are captured in a Random Indexing (RI) word space model (Kanerva et al., 2000), where semantic similarity between words is represented as proximity in high-dimensional vector space. The RI word space representation of a corpus is obtained by assigning to each unique word an initially empty, n-dimensional context vector, as well as a static, n-dimensional index vector, which contains a small number of randomly distributed non-zero elements (-1s and 1s), with the rest of the elements set to zero[1]. For each occurrence of a word in the corpus, the index vectors of the surrounding words are added to the target word's context vector. The semantic similarity between two words can then be estimated by calculating, for instance, the cosine similarity between their context vectors. A set of word space models are induced from unstructured clinical data and subsequently combined in various ways with different parameter settings (i.e., sliding window size for extracting word contexts). The models and their combinations are evaluated for their ability to map a given abbreviation to its corresponding definition. The best model achieves 42% recall. Improvement of the post-processing of candidate definitions is suggested in order to obtain enhanced performance on this task.

The estimate of word relatedness that is obtained from a word space model is purely statistical and has no linguistic knowledge. When word pairs should not only share distributional properties, but also have similar orthographic representations – as is the case for abbreviation-definition pairs – normalization procedures could be applied. Given a set of candidate definitions for a given abbreviation, the task of identifying *plausible* candidates can be viewed as a normalization problem. Petterson et al. (2013) utilize a string distance measure, Levenshtein distance (Levenshtein, 1966), in order to normalize historical spelling of words into modern spelling. Adjusting parameters, i.e., the maximum allowed distance between source and target, according to observed distances between known word pairs of historical and modern spelling, gives a normalization accuracy of 77%. In addition to using a Levenshtein distance weighting factor of 1, they experiment with context free and context-sensitive weights for frequently occurring edits between word pairs in a training corpus. The context-free weights are calculated on the basis of one-to-one standard edits involving two characters; in this setting the normalization accuracy is increased to 78.7%. Frequently occurring edits that involve more than two characters, e.g., substituting two characters for one, serve as the basis for calculating context-sensitive weights and gives a normalization accuracy of 79.1%. Similar ideas are here applied to abbreviation expansion by utilizing a normalization procedure for candidate expansion selection.

3 Method

The current study aims to replicate and extend a subset of the experiments conducted by Henriksson et al. (2012), namely those that concern the abbreviation expansion task. This includes the various word space combinations and the parameter optimization. The evaluation procedure is similar to the one described in (Henriksson et al., 2012). The current study, however, focuses on post-processing of the semantically related words by introducing a filter and a normalization procedure in an attempt to improve performance. An overview of the approach is depicted in Figure 1.

Abbreviation expansion can be viewed as a two-step procedure, where the first step involves detection, or extraction, of abbreviations, and the second step involves identifying plausible expansions. Here, the first step is achieved by extracting abbreviations from a clinical corpus with clinical abbreviation detection software and using a list of known medical abbreviations. The second step is performed by first extracting a set of semantically

[1]Generating sparse vectors of a sufficiently high dimensionality in this manner ensures that the index vectors will be *nearly* orthogonal.

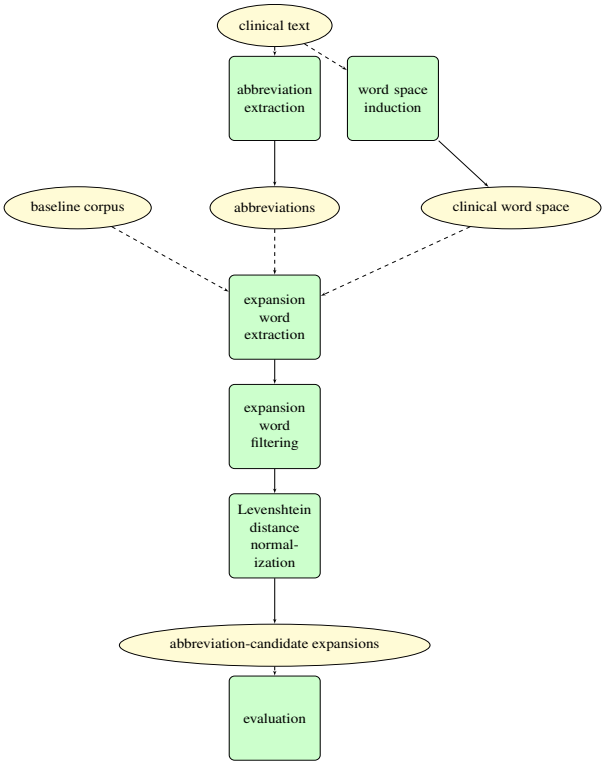

Figure 1: The abbreviation expansion process of the current study.

similar words for each abbreviation and treating these as initial expansions. More plausible expansions of each abbreviation are then obtained by filtering the expansion words and applying a normalization procedure.

3.1 Data

3.1.1 Corpora

Four corpora are used in the experiments: two clinical corpora, a medical (non-clinical) corpus and a general Swedish corpus (Table 1).

The clinical corpora are subsets of the Stockholm EPR Corpus (Dalianis et al., 2009), comprising health records for over one million patients from 512 clinical units in the Stockholm region over a five-year period (2006-2010)[2]. One of the clinical corpora contains records from various clinical units, for the first five months of 2008, henceforth referred to as SEPR, and the other contains radiology examination reports, produced in 2009 and 2010, the Stockholm EPR X-ray Corpus (Kvist and Velupillai, 2013) henceforth referred to as SEPR-X. The clinical corpora were lemmatized using Granska (Knutsson et al., 2003).

The experiments in the current study also include a medical corpus. The electronic editions of Läkartidningen (Journal of the Swedish Medical Association), with issues from 1996 to 2010, have been compiled into a corpus (Kokkinakis, 2012), here referred to as LTK.

To compare the medical texts to general Swedish, the third version of the Stockholm Umeå Corpus (SUC 3.0) (Källgren, 1998) is used. It is a balanced corpus and consists of written Swedish texts from the early 1990's from various genres.

Corpus	#Tokens	#Types	#Lemmas
SEPR	109,663,052	853,341	431,932
SEPR-X	20,290,064	200,703	162,387
LTK	24,406,549	551,456	498,811
SUC	1,166,593	97,124	65,268

Table 1: Statistical descriptions of the corpora

3.1.2 Reference standards

A list of medical abbreviation-definition pairs is used as test data and treated as the reference standard in the evaluation. The list is derived from Cederblom (2005) and comprises 6384 unique abbreviations from patient records, referrals and scientific articles. To increase the size of the test data, the 40 most frequent abbreviations are extracted by a heuristics-based clinical abbreviation detection tool called SCAN (Isenius et al., 2012). A domain expert validated these abbreviations and manually provided the correct expansion(s).

An inherent property of word space models is that they model semantic relationships between unigrams. There are, however, abbreviations that expand into multiword expressions. Ongoing research on modeling semantic composition with word space models exists, but, in the current study abbreviations that expanded to multiword definitions were simply removed from the test data set. The two sets of abbreviation-expansion pairs were merged into a single test set, containing 1231 unique entries in total.

In order to obtain statistically reliable semantic relations in the word space, the terms of interest must be sufficiently frequent in the data. As a result, only abbreviation-expansion pairs with frequencies over 50 in SEPR and SEPR-X, respectively, were included in each test set. The SEPR test set contains 328 entries and the SEPR-X test

[2]This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnamnden i Stockholm), permission number 2012/2028-31/5

set contains 211 entries. Each of the two test data sets is split into a development set (80%) for model selection, and a test set (20%) for final performance estimation.

3.2 Expansion word extraction

For the experiments where semantically related words were used for extraction of expansion words, the top 100 most correlated words for each of the abbreviations were retrieved from each of the word space model configurations that achieved the best results in the parameter optimization experiments.

The optimal parameter settings of a word space vary with the task and data at hand. It has been shown that when modeling paradigmatic (e.g., synonymous) relations in word spaces, a fairly small context window size is preferable (Sahlgren, 2006). Following the best results of Henriksson et al. (2012), we experiment with window sizes of 1+1, 2+2, and 4+4.

Two word space algorithms are explored: Random Indexing (RI), to retrieve the words that occur in a similar context as the query term, and Random Permutation (RP), which also incorporates word order information when accumulating the context vectors (Sahlgren et al., 2008). In order to exploit the advantages of both algorithms, and to combine models with different parameter settings, RI and RP model combinations are also evaluated. The models and their combinations are:

- Random Indexing (RI): words with a contextually high similarity are returned; word order within the context window is ignored.
- Random Permutation (RP): words that are contextually similar and used in the same relative positions are returned; these are more likely to share grammatical properties.
- RP-filtered RI candidates (RI_RP): returns the top ten terms in the RI model that are among the top thirty terms in the RP model.
- RI-filtered RP candidates (RP_RI): returns the top ten terms in the RP model that are among the top thirty terms in the RI model.
- RI and RP combination of similarity scores (RI+RP): sums the cosine similarity scores from the two models for each candidate term and returns the candidates with the highest aggregate score.

All models are induced with three different context window sizes for the two clinical corpora, SEPR and SEPR-X. For each corpus, two variants are used for word space induction, one where stop words are removed and one where stop words are retained. All word spaces are induced with a dimensionality of 1000.

For parameter optimization and model selection, the models and model combinations are queried for semantically similar words. For each of the abbreviations in the development set, the ten most similar words are retrieved. Recall is computed with regard to this list of candidate words, whether the correct expansion is among these ten candidates. Since the size of the test data is rather limited, 3-fold cross validation is performed on the development set for the parameter optimization experiments. For both SEPR and SEPR-X development sets, a combination of a RI model with a context window size of 4+4 and a RP model with 4+4 context window size in the summing similarity scores setting were among the most successful with recall scores of 0.25 for SEPR and 0.17 for SEPR-X.

3.3 Filtering expansion words

Given the expansion words, extracted from clinical word spaces or baseline corpora (the baselines are more thoroughly accounted for in 3.5), a filter was applied in order to generate candidate expansions. The filter was defined as a set of requirements, which had to be met in order for the expansion word to be extracted as a candidate expansion. The requirements were that the intitial letter of the abbreviation and expansion word had to be identical. All the letters of the abbreviation also had to be present in the expansion word in the same order.

String length difference was also a part of the requirements: the expansion word had to be at least one character longer than the abbreviation. In order to define an upper bound for expansion token length, string length differences of the SEPR and SEPR-X development sets were obtained. The distribution of string length differences for abbreviation-expansion pairs in the SEPR development set ranged from 1 to 21 characters. If a maximum string length difference of 14 was allowed, 95.2% of the abbreviation-expansion pairs were covered. As for the string length differences in the SEPR-X development set, the distribution ranged from 1 to 21 characters. If a string length difference of up to and including 14 characters was allowed, 96.3% of the abbreviation-expansion pairs were covered. Thus, a maximum difference

in string length of 14 was also required for the expansion word to be extracted as a candidate expansion.

3.4 Levenshtein distance normalization

Given the set of filtered candidate expansions for the abbreviations, choosing the correct one can be seen as a normalization problem. The goal is to map a source word to a target word, similarly to for instance methods for spelling correction. The target word is chosen from a list of words, and the choice is based on the distance between the source and the target where a small distance implies high plausibility. However, we cannot adopt the same assumptions as for the problem of spelling correction, where the most common distance between a source word and the correct target word is 1 (Kukich, 1992). Intuitively, we can expect that there are abbreviations that expand to words within a larger distance than 1. It would seem somewhat useless to abbreviate words by one character only, although it is not entirely improbable.

Similarly to measuring the string length difference in order to define an upper bound for filtering candidate expansions, the Levenshtein distances for abbreviation-expansion pairs in the development sets were obtained.

For the SEPR and SEPR-X development sets, allowing a Levenshtein distance up to and including 14 covers 97.8% and 96.6% of the abbreviation-expansion pairs, as shown in Table 2.

Given the filtered candidate expansions, the Levenshtein distance for the abbreviation and each of the candidate expansions were computed. For each one of the candidate expansions, the Levenshtein distance beween the entry and the abbreviation was associated with the entry. The resulting list was sorted in ascending order according to Levenshtein distance.

Going through the candidate expansion list, if the Levenshtein distance was less than or identical to the upper bound for Levenshtein distance (14), the candidate expansion was added to the expansion list that was subsequently used in the evaluation. In the Levenshtein distance normalization experiments, a combination of semantically related words and words from LTK was used. When compiling the expansion list, semantically related words were prioritized. This implied that word space candidate expansion would occupy the top positions in the expansion list, in ascending order according to Levenshtein distance. The size of the list was restricted to ten, and the remaining positions, if there were any, were populated by LTK candidate expansions in ascending order according to Levenshtein distance to the abbreviation. If there were more than one candidate expansion at a specific Levenshtein distance, ranking of these was randomized.

	SEPR	SEPR	SEPR-X	SEPR-X
LD	Avg %	SDev	Avg %	SDev
1	1	0.3	0.4	0.2
2	4.6	0.4	5	0.6
3	13	1.2	14.7	1.3
4	12.2	1	15.1	0.6
5	12.7	1.3	14.5	2.2
6	12.7	0.8	12.9	0.9
7	8.4	0.7	7.8	0.3
8	10.4	1.5	9.8	2
9	5.7	0.7	4.9	0.5
10	4.1	0.7	2.9	0.3
11	3	0.5	2.6	0.4
12	3	0.6	2.6	0.4
13	3.8	5.5	1.3	0.5
14	3.5	1.1	2.2	0.8
15	1.3	0.5	1.3	0.5
16	1.6	0.4	0.4	0.2
17	0.2	0.1		
18	0.8	0.3	1	0.1
20	0.2	0.1		
21	0.2	0.1	0.5	0

Table 2: Levenshtein distance distribution for abbreviation-expansion pairs. Average proportion over 5 folds at each Levensthein distance with standard deviation (SDev) in SEPR and SEPR-X development sets.

3.5 Evaluation

The evaluation procedure of the abbreviation expansion implied assessing the ability of finding the correct expansions for abbreviations. In order to evaluate the performance gain of using semantic similarity to produce the list of candidate expansions over using the filtering and normalization procedure alone, a baseline was created. For the baseline, expansion words were instead extracted from the baseline corpora, the corpus of general Swedish SUC 3.0 and the medical corpus LTK. A list of all the lemma forms from each baseline

corpus (separately) was provided for each abbreviation as initial expansion words. The filter and normalization procedure was then applied to these expansion words.

The reference standard contained abbreviation-expansion pairs, as described in 3.1.2. If any of the correct expansions (some of the abbreviations had multiple correct expansions) was present in the expansion list provided for each abbreviation in the test set, this was regarded as a true positive. Precision was computed with regard to the position of the correct expansion in the list and the number of expansions in the expansion list, as suggested in Henriksson (2013). For an abbreviation that expanded to one word only, this implied that the expansion list besides holding the correct expansion, also contained nine incorrect expansions, which was taken into account when computing precision. The list size was static: ten expansions were provided for each abbreviation, and this resulted in an overall low precision. Few of the abbreviations in the development set expanded to more than one word, giving a precision of 0.17-0.18 for all experiments.

Results of baseline abbreviation expansion in the development sets are given in table 3. Recall is given as an average of 5 folds, as cross validation was performed. The baseline achieves overall low recall, with the lowest score of 0.08 for the SEPR-X development set using SUC for candidate expansion extraction. The rest of the recall results are around 0.11.

Corpus	SEPR	SEPR	SEPR-X	SEPR-X
	Recall	SDev	Recall	SDev
SUC	0.10	0.05	0.08	0.06
LTK	0.11	0.06	0.11	0.11

Table 3: Baseline average recall for SEPR and SEPR-X development sets.

Results from abbreviation expansion using semantically related words with filtering and normalization to refine the selection of expansions on SEPR and SEPR-X development sets are shown in Table 4. Recall is given as an average of 5 folds, as cross validation was performed. The semantically related words are extracted from the word space model configuration that had the top recall scores in the parameter optimization experiments described in 3.2, namely the combination of an RI model and an RP model both with 4+4 context window sizes. Recall is increased by 14 percentage points for SEPR and 20 percentage points for SEPR-X when applying filtering and normalization to the semantically related words.

SEPR	SEPR	SEPR-X	SEPR-X
Recall	SDev	Recall	SDev
0.39	0.05	0.37	0.1

Table 4: Abbreviation expansion results for SEPR and SEPR-X development sets using the best model from parameter optimization experiments (RI.4+4+RP.4+4).

4 Results

4.1 Expansion word extraction

The models and model combinations that had the best recall scores in the word space parameter optimization were also evaluated on the test set. The models that had top recall scores in 3.2 achieved 0.2 and 0.18 for SEPR and SEPR-X test sets respectively, compared to 0.25 and 0.17 in the word space parameter optimization.

4.2 Filtering expansion words and Levenshtein normalization

Abbreviation expansion with filtering and normalization was evaluated on the SEPR and SEPR-X test sets. The results are summarized in Table 5.

	SEPR	SEPR-X
SUC	0.09	0.16
LTK	0.08	0.14
Expansion word extraction	0.20	0.18
Filtering and normalization	0.38	0.40

Table 5: SEPR and SEPR-X test set results in abbreviation expansion.

Baseline recall scores were 0.09 and 0.08 for SUC and LTK respectively, showing a lower score for LTK compared to the results on the SEPR development set. For abbreviation expansion (with filtering and normalization) using semantically related words in combination with LTK, the best recall score was 0.38 for the SEPR test set, compared to 0.39 for the same model evaluated on the SEPR development set. Compared to the results of using semantically related words only (expansion word extraction), recall increased by 18 percent-

age points for the same model when filtering and normalization was applied.

Evaluation on the SEPR-X test set gave higher recall scores for both baseline corpora compared to the baseline results for the SEPR-X development set: the SUC result increased by 8 percentage points for recall. For LTK, there was an increase in recall of 3 percentage points. For the SEPR-X test set, recall increased by 22 percentage points when filtering and normalization was applied to semantically related words extracted from the best model configuration.

In comparison to the results of Henriksson et al (2012), where recall of the best model is 0.31 without and 0.42 with post-processing of the expansion words for word spaces induced from the data set (i.e., an increase in recall by 11 percentage points), the filtering and normalization procedure for expansion words of the current study yielded an increase by 18 percentage points.

5 Discussion

The filter combined with the Levenshtein normalisation procedure to refine candidate expansion selection showed a slight improvement compared to using post-processing, although the normalization procedure should be elaborated in order to be able to confidently claim that Levenshtein distance normalization is a better approach to expansion candidate selection. A suggestion for future work is to introduce weights based on frequently occurring edits between abbreviations and expansions and to apply these in abbreviation normalization.

The approach presented in this study is limited to abbreviations that translate into *one* full length word. Future research should include handling multiword expressions, not only unigrams, in order to process acronyms and initialisms.

Recall of the development sets in the word space parameter optimization experiments showed higher scores for SEPR (0.25) compared to SEPR-X (0.17). An explanation to this could be that the amount of data preprocessing done prior to word space induction might have varied, in terms of excluding sentences with little or no clinical content. This will of course affect word space co-occurrence information, as word context is accumulated without taking sentence boundaries into account.

The lemmatization of the clinical text used for word space induction left some words in their original form, causing test data and semantically related words to be morphologically discrepant. Lemmatization adapted to clinical text might have improved results. Spelling errors were also frequent in the clinical text, and abbreviations were sometimes normalized into a misspelled variant of the correct expansion. In the future, spelling correction could be added and combined with abbreviation expansion.

The impact that this apporach to abbreviation expansion might have on readability of clinical texts should also be assessed by means of an extrinsic evaluation, a matter to be pursued in future research.

6 Conclusions

We presented automatic expansion of abbreviations consisting of unigram full-length words in clinical texts. We applied a distributional semantic approach by using word space models and combined this with Levenshtein distance measures to choose the correct candidate among the semantically related words. The results show that the correct expansion of the abbreviation can be found in 40% of the cases, an improvement by 24 percentage points compared to the baseline (0.16) and an increase by 22 percentage points compared to using word space models alone (0.18). Applying Levenshtein distance to refine the selection of semantically related candidate expansions yields a total recall of 0.38 and 0.40 for radiology reports and medical health records, respectively.

Acknowledgments

The study was partly funded by the Vårdal Fundation and supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (ref. no. IIS11-0053) at Stockholm University, Sweden. The authors would also like to direct thanks to the reviewers for valuable comments.

References

M. Adnan, J. Warren, and M. Orr. 2010. Assessing text characteristics of electronic discharge summaries and their implications for patient readability. In *Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management-Volume 108*, pages 77–84. Australian Computer Society, Inc.

H. Allvin. 2010. Patientjournalen som genre: En text- och genreanalys om patientjournalers relation till patientdatalagen. Master's thesis, Stockholm University.

H. Ao and T. Takagi. 2005. ALICE: an algorithm to extract abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, 12(5):576–586.

S. Cederblom. 2005. *Medicinska förkortningar och akronymer (In Swedish)*. Studentlitteratur.

J.T. Chang, H. Schütze, and R.B. Altman. 2002. Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9:612–620.

T. Cohen and D. Widdows. 2009. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405.

H. Dalianis, M. Hassel, and S. Velupillai. 2009. The Stockholm EPR Corpus – Characteristics and some initial findings. In *Proceedings of the 14th International Symposium on Health Information Management Research*, pages 243–249.

D. Dannélls. 2006. Automatic acronym recognition. In *Proceedings of the 11th conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 167–170.

N. Elhadad. 2006. *User-sensitive text summarization: Application to the medical domain*. Ph.D. thesis, Columbia University.

S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. 2005. Resolving abbreviations to their senses in MEDLINE. *Bioinformatics*, 21(18):3658–3664, September.

Z.S. Harris. 1954. Distributional structure. *Word*, 10:146–162.

A. Henriksson, H. Moen, M. Skeppstedt, A. Eklund, V. Daudaravicius, and M. Hassel. 2012. Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models. In *Proceedings of Semantic Mining in Biomedicine (SMBM 2012)*, pages 10–17.

A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius, and M. Duneld. 2014. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(6).

A. Henriksson. 2013. *Semantic Spaces of Clinical Text: Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records*. Licentiate thesis, Department of Computer and Systems Sciences, Stockholm University.

N. Isenius, S. Velupillai, and M. Kvist. 2012. Initial Results in the Development of SCAN: a Swedish Clinical Abbreviation Normalizer. In *Proceedings of the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis (CLEFeHealth2012)*.

G. Källgren. 1998. Documentation of the Stockholm-Umeå corpus. *Department of Linguistics, Stockholm University*.

P. Kanerva, J. Kristoferson, and A. Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the cognitive science society*, page 1036.

A. Keselman, L. Slaughter, C. Arnott-Smith, H. Kim, G. Divita, A. Browne, C. Tsai, and Q. Zeng-Treitler. 2007. Towards consumer-friendly PHRs: patients experience with reviewing their health records. In *AMIA Annual Symposium Proceedings*, volume 2007, pages 399–403.

O. Knutsson, J. Bigert, and V. Kann. 2003. A robust shallow parser for Swedish. In *Proceedings of Nodalida*.

D. Kokkinakis. 2012. The Journal of the Swedish Medical Association-a Corpus Resource for Biomedical Text Mining in Swedish. In *Proceedings of Third Workshop on Building and Evaluating Resources for Biomedical Text Mining Workshop Programme*, page 40.

K. Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439.

M. Kvist and S. Velupillai. 2013. Professional Language in Swedish Radiology Reports – Characterization for Patient-Adapted Text Simplification. In *Scandinavian Conference on Health Informatics 2013*, pages 55–59.

V.I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707.

D. Movshovitz-Attias and W.W. Cohen. 2012. Alignment-HMM-based Extraction of Abbreviations from Biomedical Text. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012)*, pages 47–55.

N. Nakatsu, Y. Kambayashi, and S. Yajima. 1982. A longest common subsequence algorithm suitable for similar text strings. *Acta Informatica*, 18(2):171–179.

Y. Park and R.J. Byrd. 2001. Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the 2001 conference on empirical methods in natural language processing*, pages 126–133.

E. Pettersson, B. Megyesi, and J. Nivre. 2013. Normalisation of historical text using context-sensitive weighted levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 163–179.

R.E. Rudd, B.A. Moeykens, and T.C. Colton. 1999. Health and literacy: a review of medical and public health literature. *Office of Educational Research and Improvement*.

M. Sahlgren, A. Holst, and P. 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.

M. Sahlgren. 2006. *The Word-space model*. Ph.D. thesis, Stockholm University.

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

M. Skeppstedt. 2012. *From Disorder to Order: Extracting clinical findings from unstructured text*. Licentiate thesis, Department of Computer and Systems Sciences, Stockholm University.

K. Taghva and J. Gilbreth. 1999. Recognizing acronyms and their definitions. *International Journal on Document Analysis and Recognition*, 1(4):191–198.

H. Yu, G. Hripcsak, and C. Friedman. 2002. Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association*, 9(3):262–272.