Towards a learning approach for abbreviation detection and resolution

Klaar Vanopstal[1,2], **Bart Desmet**[1,2] **and Véronique Hoste**[1,2]

[1]LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
klaar.vanopstal, bart.desmet, veronique.hoste@hogent.be
[2]Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium

Abstract

The explosion of biomedical literature and with it the -uncontrolled- creation of abbreviations presents some special challenges for both human readers and computer applications. We developed an annotated corpus of Dutch medical text, and experimented with two approaches to abbreviation detection and resolution. Our corpus is composed of abstracts from two medical journals from the Low Countries in which approximately 65 percent (NTvG) and 48 percent (TvG) of the abbreviations have a corresponding full form in the abstract. Our first approach, a pattern-based system, consists of two steps: abbreviation detection and definition matching. This system has an average F-score of 0.82 for the detection of both defined and undefined abbreviations and an average F-score of 0.77 was obtained for the definitions. For our second approach, an SVM-based classifier was used on the preprocessed data sets, leading to an average F-score of 0.93 for the abbreviations; for the definitions an average F-score of 0.82 was obtained.

1. Introduction

With the explosion of biomedical information, the number of biomedical abbreviations is growing rapidly. As there are no rules for the formation of new abbreviations, their detection and association to the correct full form becomes increasingly difficult. Abbreviation detection is especially useful for language technology applications like information retrieval (Byrd et al., 1994) and extraction (Maynard and Ananiadou, 1999; Roark and Charniak, 1998; Liu et al., 2002), NER and anaphora resolution. Yu et al. (2002b) claim that exploiting abbreviations in IR systems increases the number of relevant documents retrieved, and Friedman et al. (2001) argue that not handling abbreviations in NLP is a major source of errors.

In this paper, we describe the compilation of a monolingual corpus of Dutch medical texts which can be used as a gold standard for the detection of abbreviations. Chang and Schütze (2006) mention the lack of a suitable standard for English as one of the factors that hamper the creation and evaluation of efficient abbreviation detection systems. In section 3, we describe the creation of a Dutch gold standard for abbreviation detection and resolution and the guidelines used for the annotation of the corpus. Section 4 presents some challenges for the automatic detection of abbreviation patterns and in section 5, we refer to related work and we present our approach. Finally, in section 6, we present some conclusions and describe our future work.

2. Creation of a gold standard

2.1. Corpus

As the construction of a gold standard is both time-consuming and expensive, such inventories are scarce, and even more so for a niche language like Dutch. Chang and Schütze (2006) list four criteria for a solid gold standard: size, breadth, accuracy and common use in the community. A bigger corpus generally leads to a larger coverage and thus to a higher accuracy. As we are focusing on a specific domain, namely the medical domain, the second criterium, i.e. breadth of the corpus, is of lesser importance to our research.

There are two popular gold standards for English abbreviation detection: the AbbRE gold standard and the Medstract acronym gold standard. The corpus used for the evaluation of the AbbRE algorithm (Yu et al., 2002a) is claimed to be reliable, as it was annotated by three experts in the field. However, its size is next to negligible: it consists of only 10 articles. Medstract, a publicly available and commonly used standard, is somewhat larger but seems to be less reliable (Chang and Schütze, 2006). For Dutch, no such resources are available.

For the construction of our corpus, we randomly selected abstracts from the Dutch journal *Nederlands Tijdschrift voor Geneeskunde*[1] (NTvG) and the Belgian journal *Belgisch Tijdschrift voor Geneeskunde*[2](TvG), two important sources of information for Dutch-speaking experts in the biomedical sector in the Netherlands and in Belgium. This resulted in a corpus of 66,739 words, 29,978 (100 abstracts) of which are from the NTvG, and 36,757 (256 abstracts) from the TvG. The corpus was tokenized before the annotation.

2.2. Annotation guidelines

In order to be able to mark all abbreviations in a text, we needed to define what we consider abbreviations, and what distinguishes them from acronyms, clipping and other short forms. The definitions found in the literature vary, but generally, abbreviations are described as *a shortened form of a word or phrase* (Chang and Schütze, 2006).

Liu et al. (2001) distinguish six types of abbreviations:

1. Truncating the end, e.g. *adm* for *administration* (or *administrator*), also called clipping (Bloom, 2000)

[1] http://www.ntvg.nl

[2] http://poj.peeters-leuven.be/content.php?url=journal&journal_code=TVG

2. First letter initialization, or acronyms, e.g. *AAA* for *abdominal aortic aneurysm*

3. Opening letter initialization, e.g. *HeLa* for *Henrietta Lacks*

4. Syllabic initialization, e.g. *BZD* for *benzodiazepine*

5. Combination initialization, e.g. *e-mail* for *electronic mail*

6. Substitution initialization e.g. *ASD I* for *Primum atrial septal defect*; *Fe* for *iron*.

Liu et al. (2002) add a seventh type of abbreviation, i.e. the chemical abbreviation, which is a combination of letters and numbers.

CXCR4	chemokinereceptor fusine (EN: chemokine receptor fusin)
CTLA4-Ig	cytotoxisch T-lymfocyt-A4-Ig (EN: cytotoxic T-lymphocyte-associated antigen-4 Ig)

We used eight labels for the annotation of our corpus:

1. **ABBR**: Dutch abbreviations which have a full form in their local context.

    (1) Hoge-resolutie-computertomografie (**HRCT**) kan een belangrijke rol spelen in [...]. (EN: High resolution computed tomography (HRCT) can play an important role in [...].)

2. **ABBR_DE**: Dutch abbreviations which have a full form somewhere (DE = defined elsewhere) in the same abstract, but not in their local context.

    (2) Het chronische-vermoeidheidssyndroom (*chronic fatigue syndrome* , CFS) is [...]. Tegenwoordig wordt meer belang gehecht aan de mogelijke rol van het centrale zenuwstelsel (CZS) in de pathofysiologie van het **CFS** . (EN: The chronic fatigue syndrome (CFS) is [...]. The possible role of the central nerve system (CNS) is increasingly emphasized in the pathophysiology of CFS.)

3. **DEF**: Dutch full forms which define an abbreviation in their local context.

    (3) Recent onderzoek betreffende de atrofie van de **mediale temporale kwab** (MTL), gaf bemoedigende resultaten voor de aanvullende diagnostiek van de **ziekte van Alzheimer** (AD)[...]. (EN: Recent studies about the atrophy of the medial temporal lobe (MTL) showed optimistic results for the supplementary diagnostics of Alzheimer's disease (AD)[...].)

    (4) **Gesystematiseerde lupus erythematosus** (SLE) is hèt prototype van [...]. (EN: Systematized lupus erythematosus (SLE) is the prototype of [...])

    As can be observed in example 3, this Dutch definition can also refer to an English abbreviation.

4. **ABBR_IN_COMP**: Abbreviations which are part of a compound word and which do not have a definition in the abstract.

    (5) De fysiopathologische veranderingen zijn het gevolg van het tegenwerkend effect van een primair verhoogde **ADH**-secretie en een secundair gestegen **ANF**-vrijzetting. (EN: These physiopathological changes are due to the countereffect of a primary inappropriate of ADH secretion and a secondary increased ANF secretion.)

    (6) Men kan dez6 systemen in twee grote categorieën onderverdelen : de **Ca**-transport- **ATP**asen, die [...]. (EN: These systems are divided into 2 large categories: the Ca-Transport ATPases, which [...]

    Usually, abbreviations in Dutch compounds are separated from the other part of the compound with a hyphen. In some cases (see example 21), no such separation is used (ATPasen).

5. **ABBR_IN_COMP_DE**: Abbreviations which are part of a compound word and which have a full form or definition in the abstract.

    (7) [...] ernstige *reumatoïde artritis* (RA)-vasculitis. Bij de ziekte van Wegener en **RA**-vasculitis [...]. (EN: [...] severe rheumatoid arthritis (RA) vasculitis. Wegener's disease and RA vasculitis [...])

6. **ABBR_NO_DEF**: These are abbreviations which have no definition or full form in their local context. However, the term's full form occurs somewhere else in the text.

    (8) [...] orale **5-FU** prodrugs (capecitabine, **UFT**), ethynyluracil en [...]. (EN: [...] oral 5-FU prodrugs (capecitabine, UFT), ethynyluracil and [...].)

7. **ABBR_EN**: An English abbreviation, with either a Dutch or an English definition in its local context.

    (9) [...] gecorreleerd aan de score op de Mini Mental State Examination (**MMSE**). (EN: correlated with the Mini Mental State Examination (MMSE) score.)

    (10) Endogeen stikstofmonoxide (**NO**) speelt een belangrijke rol in [...] (EN: Endogenous nitric oxide (NO) plays an important role in [...])

8. **DEF_EN**: An English definition or full form which accompanies an English abbreviation.

    (11) [...] hebben hogere serum-leptinewaarden in vergelijking met mannen met eenzelfde "**body mass index**" (BMI). (EN: have higher serum leptin values when compared to men with the same "body mass index" (BMI).)

The kappa score (Carletta, 1996), which is an indication of the agreement between the annotators of the corpus, is 0.89. Table 1 shows the frequency of the labels used in our annotations.

	NTvG	TvG
ABBR	117	115
DEF	138	146
ABBR_DE	309	182
ABBR_EN	62	54
DEF_EN	45	30
ABBR_NO_DEF	279	235
ABBR_IN_COMP	72	181
ABBR_IN_COMP_DE	170	40

Table 1: Frequency of the labels used in our annotations

Table 2 below shows the proportion of abbreviations in the corpus: 3.36% (1,009 abbreviations out of 29,978 words) in the NTvG and 2.19% (807 abbreviations out of 36,757 words) in the Tvg. A total of 17.74% of the abbreviations in the Dutch corpus are defined abbreviations (Yu et al., 2002a) (i.e. abbreviations with a definition in English or Dutch in their local context), compared to 20.94% in the Belgian corpus. This number includes anaphoric and cataphoric patterns of Dutch or English abbreviations with a Dutch or English definition in their local context. The undefined abbreviations include Dutch abbreviations without a definition and abbreviations in compounds which have not been defined elsewhere in the abstract. When we broaden the scope to abbreviations which have a definition in the abstract (i.e. not only in the local context), the proportion of defined abbreviations increases to 65.21% for the Dutch and to 48.45% for the Belgian corpus (Table 2). This means that, for both corpora, between 45% and 52% of the abbreviations remain unresolved.

	NTvG	TvG
abbr	3.36%	2.19%
defined abbr	17.74%	20.94 %
abbr defined in broader context	47.47%	27.50%
abbr defined in local or broader context	65.21%	48.45%

Table 2: Proportion of abbreviations and defined abbreviations in the corpus

3. Challenges for the automatic detection of abbreviations

Some "irregularities" in the formation or patterns of abbreviations can form a challenge to the automatic detection of abbreviations and their definitions.

First of all, many of the English abbreviations in Dutch texts which have a Dutch definition cannot be matched to the initial letters of the definition.

(12) HAART **k**rachtige **a**ntiretrovirale **t**herapie
(EN: highly active anti-retroviral therapy)

Another challenge for the system will be the distinction between ordinary parenthetical patterns and parenthetical patterns which include an abbreviation and a definition, e.g.

(13) gunstige uitkomst (**score 5**)
(EN: positive result (score 5))

Due to its different compounding rules, Dutch has more compounds which are written in one single word than English. Consequently, the letters in abbreviations have to be aligned to syllables or word parts rather than initial letters.

(14) CVS **c**hronische-**v**ermoeidheid**s**yndroom
(EN: chronic fatigue syndrome)

Not all words are included in the abbreviation. Usually, function words are left out. In these cases, POS tagging or stop word lists can help the resolution process or the system can allow for words to be in the definition which are not initialized in the abbreviation (Liu et al., 2002).

(15) ADL **a**ctiviteiten **v**an **h**et **d**agelijks **l**even
(EN: daily life activities)

According to Liu et al. (Liu et al., 2001), the average number of different full forms for all abbreviations with six characters or less is 2.28. The longer the abbreviation, the less ambiguous it will be. This is especially challenging for the resolution of undefined abbreviations, i.e. abbreviations which have no definition in their local context, e.g.

(16) PMR **p**sycho**m**otore **r**etardatie OR **p**oly**m**yalgia **r**heumatica
(EN: psychomotor retardation OR polymyalgia rheumatica)

4. Automatic abbreviation detection and resolution

4.1. Related research

Different methodologies have been applied to the detection and resolution of abbreviations. The most popular approach relies on the use of heuristics to detect patterns of uppercase words consisting of a limited number of letters, which occur in predefined constructions. Taghva and Gilbreth's (1999) system, for example, identifies words of 3-10 uppercase letters as candidate abbreviations. An algorithm matches the letters of the abbreviation against the initial letters of a candidate definition, which is situated in a window of 2*N (where N stands for the number of letters in the abbreviation) words next to the abbreviation. The authors report recall and precision rates of 86% and 98% for definitions respectively and 93% and 98% for the abbreviations. The Stanford Medical Abbreviation Method (Chang and Schütze, 2006) uses 3*N instead of 2*N. This is a hybrid system, using both rule-based and machine learning (cf. infra) techniques. The ARGH[3] (Acronym Resolving General Heuristics) method (Wren and Garner, 2002) uses a set of heuristics and refinement rules to identify the boundaries of acronym-definition pairs. It treats parenthetical constructions consisting of one word as candidate acronyms and attempts to match each acronym letter to letters within the words immediately to the left of it. Conversely, parenthetical patterns consisting of multiple words are considered

[3] http://invention.swmed.edu/argh/

as candidate definitions. The advantage of this system is that it is able to recognize word patterns that are not in the same order as the acronym letters (e.g. *ACMV-NOg* for ***A**frican **c**assava **m**osaic **v**irus isolate **o**riginating from **N**igeria*). The system achieves an estimated rate of 96.5% precision and 93% recall when tested on over 12 million MEDLINE records. Another example of a heuristics-based approach is presented by Park and Byrd (2001) who combine text markers (e.g. “(..)”, “[...]” and “=”) and linguistic cues (e.g. “short”, “or” and “stand”) with pattern-based recognition. In addition to heuristics, natural language processing tools can be used to refine the search space of the definition. Pustojevski et al. (2001), for example, use part-of-speech information and only consider noun phrases as candidate definitions.

As an alternative to the heuristics-based approaches, a machine learning approach can be adopted. Chang et al. (2002) present a supervised learning algorithm for acronym identification, which only searches for “definition (acronym)” patterns, as this is the most frequently used pattern. Chang et al. use nine features to score the abbreviations (such as percentage of letters aligned at the beginning of a word, number of definition words that were skipped in the alignment, percentage of letters aligned on syllable boundary etc.). They report a 83% recall and 80% precision on the Medstract corpus composed of MEDLINE abstracts (Pustejovsky et al., 2002).

4.2. Pattern-based approach

Our pattern-based approach can be divided into two steps, as in most of the systems mentioned above: abbreviation detection and definition matching, based on some predefined patterns. In the first step, we tried to detect constructions which consist of capital letters, or combinations of capital letters with one to three lowercased letters and/or numbers. Parenthetical constructions and other text markers (e.g. “=”, “ ” ”, “ ’ ”) (Park and Byrd, 2001) serve as indicators for abbreviation-definition combinations and are used to match the abbreviations to their full forms in the text in the second step. In the abbreviation detection stage, abbreviations are printed with their candidate definition, which is situated within a range of 3*N (Chang and Schütze, 2006) words preceding or following the abbreviation.

In the second step, i.e. filtering the definitions from the list of candidate definitions, we faced some of the challenges described in section 4: combinations of English abbreviations with Dutch definitions, which could not be matched on the basis of their initial letters, parenthetical patterns which did not contain an abbreviation-definition combination, the different compounding rules in Dutch, which made matching to the definition more difficult. We matched the first letter in the abbreviations against the words in the candidate definition and considered the matching word and the following words -until the end of the 3*N sequence- as a definition. In our future research, we intend to apply decompounding techniques to match all the letters in the abbreviations to word and/or word-part boundaries. The results of both the abbreviation detection and abbreviation resolution are presented in table 3.

Abbreviations			
	precision	recall	FB1
TvG	83.89	78.64	81.18
NTvG	82.05	83.07	82.56
Definitions			
	precision	recall	FB1
TvG	74.49	83.36	78.68
NTvG	68.03	85.50	75.77

Table 3: Results of the patern-based approach.

4.3. Learning approach

In addition to the pattern-based approach, we experimented with a machine learning approach for the detection and resolution of abbreviations in Dutch texts. Whereas the pattern-based approach relies on a set of pre-defined patterns which are applied to raw text, in the machine learning approach we start from a corpus which has undergone the following preprocessing steps, viz. POS tagging and NP chunking performed by a memory-based shallow parser (Daelemans and van den Bosch, 2005).

The following information was encoded in the feature vector both for the detection and resolution of abbreviations: token, part-of-speech of the token, binary features indicating whether the token is an initial, part of a URL or in sentence initial position, morphological binary features to indicate whether the token starts with a capital letter, is completely capitalized, is a roman number, contains internal capital letters, is completely lowercased, contains or is completely composed of digits, punctuation, hyphens, a feature indicating whether the token incorporates vowels, etc. Also prefix and suffix information (n = 4) was incorporated in the feature vector. A symbolic word shape feature was used to merge all information encapsulated in the morphological binary features. For the resolution of the abbreviations, a simple heuristic feature was added which matches the first letter in the detected abbreviations against the words in the local context (context of 3*N sequence).

YamCha (Kudo and Matsumoto, 2003) (version 0.33), an open source text chunker using Support Vector Machines, was used for the learning experiments. The experiments were conducted with standard parameter settings and in a ten-fold cross-validation set-up. Table 4 gives an overview of the results both for abbreviation and definition detection. For the detection of abbreviations, the SVM-based classifier yields an performance increase of 10% and more over the pattern-based approach. For the definition detection, the learning-based approach has led to a large increase of precision, leading to an overall F-score improvement of 5% over the pattern-based approach. Given the basic features which were used for this definition detection, we believe that there is certainly room for improvement for the results reported in Table 4; a more mature approach would definitely benefit from features encoding decompounding information in order to detect word-internal mappings (e.g. CVS: **c**hronische-**v**ermoeidheid**s**yndroom), features which allow for a mapping between English abbreviations and Dutch definitions, etc.

Abbreviations			
	precision	recall	FB1
TvG	95.31	92.26	93.76
NTvG	96.76	90.97	93.78
Definitions			
	precision	recall	FB1
TvG	86.92	78.18	82.32
NTvG	87.19	78.00	82.34

Table 4: Ten-fold cross-validation results of the learning experiments.

4.4. Manual error analysis

In order to better understand the errors on the abbreviation detection and resolution, we performed a manual error analysis on the output of the more error-prone pattern-based approach. This analysis can help us both to define new patterns and to develop new features for the learner. For both detection tasks, we distinguished between false positive and false negative results.

4.4.1. Abbreviation detection

In our error analysis of the abbreviation detection, we distinguished five possible causes for **false positive** results, i.e. words that were detected as abbreviations, but which had not been annotated as such:

1. Titles which are printed in capital letters:

   (17) FUNCTIONELE MRI : **HET** AFBEELDEN VAN MOTORISCHE HERSENFUNCTIES (EN: FUNCTIONAL MRI: AN IMAGE OF THE MOTOR BRAIN FUNCTIONS)

   In this example, the colon was seen as a text marker indicating the possible presence of an abbreviation. Our system detected ”HET” as an abbreviation.

2. First name initials which were erroneously detected as an abbreviation:

   (18) V. **d**e Vries

3. Roman numerals which are confused with a capitalized i, v or x:

   (19) Een coagulase-positieve Staphylococcus aureus behorende tot faaggroep **II** [...] (EN: A coagulase-positieve Staphylococcus aureus of phage group II [...])

4. Single letters which are not abbreviations:

   (20) hepatitis **A**, **B** en **C** (EN: hepatitis A, B and C)

   In this example, A, B and C are not abbreviations, but rather a type of the disease.

5. Numerals in compounds consisting of an abbreviation, a numeral and a word.

   (21) het PS-**1**-gen op chromosoom 14 (EN: PS1-1 gene on chromosome 14)

   In this case, the system was not able to decide whether “-1” belonged to the compound word “1-gen” or to the abbreviation “PS-1”.

The **false negatives**, on the other hand, were caused by the lack of rules which cover specific abbreviation formation patterns, such word-internal capital letters (22), the use of numerals and hyphens in compounds (23), single (lowercase) letter abbreviations (24), etc. Also abbreviations which do not have the orthographical characteristics of an abbreviation (25) were not detected. Most of these false negatives, however, were solved in the classification-based approach.

(22) mmHg (EN: Torr)

(23) orale **5**-FU prodrugs (EN: oral 5-FU prodrugs)

(24) leeftijd (>60 j) (EN: age (>60))

(25) 15-20 min (EN: 15-20 minutes)

4.4.2. Definition detection

The pattern-based definition detection step suffered from error percolation from the abbreviation detection, leading to both false positives (26, 27), indicated with “[]” in the examples below, and false negatives (28). Since both detection problems were handled completely independently as a binary classification task in the learning experiments, this error percolation was avoided in the classification approach.

(26) [**E**LEKTRONENBESTRALING EFFECTIEF BIJ DE BEHANDELING VAN HUIDCARCINOMEN]; **EEN** VERGELIJKING MET RÖNTGENCONTACTTHERAPIE (EN: ELECTRON IRRADIATION EFFECTIVE FOR THE TREATMENT OF SKIN CARCINOMAS; A COMPARISON WITH CONTACT X-RAY THERAPY)

(27) In dit verband werd in deel **I** (6) [**i**ngegaan] op de rol van [...] (EN: In part I (6), we examine the role of [...])

(28) [...] heeft de Werkgroep Suïcide Onderzoek Vlaanderen (**WeSOV**) [...] (EN: the Workgroup Suicide Research in Flanders (SRF) [...])

Other **false positives** in the definition detection can be attributed to:

1. Linking of the first letter of the abbreviation to another word in the sentence which starts with the same letter. The words following this mislinked word are also often erroneously labeled as parts of the definition. Since the pattern-based approach operates on raw text, function words could be detected as first word of a definition. This problem was handled in the classification-based approach by incorporating part-of-speech information.

   (29) **h**et hepatitis-A-virus (HAV) (EN: the hepatitis-A-virus (HAV))

2. Parenthetical patterns which contain an abbreviation that is not locally defined. In the example below, HLA-H and HFE are examples of a “hemochromatosegen”:

(30) Sedert de ontdekking van het hemochromatosegen (HLA-H- of HFE-gen)[...] (EN: Since the discovery of the hemochromatosis gene (HLA-H or HFE gene)[...])

False negatives could be observed in the following cases:

1. The use of function words, which cannot be linked to the letters in the abbreviation:

   (31) **op** evidentie gebaseerde zorg (EGZ) (EN: evidence-based medicine (EBM))

2. English abbreviations with Dutch definitions or expansions.

   (32) [...] voorkomen van een **overmatig waterverlies** (TEWL). (EN: prevention of **t**ransepidermal **e**xcessive **w**ater **l**oss (TEWL).)

3. Definitions which describe the concept of the abbreviation, rather than giving an expansion of it:

   (33) **lange arm van het Y-chromosoom** (Yq) (EN: long arm of the Y chromosome (Yq))

4. The use of contractions. In the example below, "therapiegebonden" is not repeated in the noun phrase "acute leukemie", although the "t" in "t-AL" refers to that same adjective.

   (34) De incidentie van **therapiegebonden** secundaire myelodysplasie (**t** - MDS) en acute leukemie (**t** - AL). (EN: the incidence of therapy-related secondary myelodysplasia (t-MDS) and acute leukemia (t-AL).)

5. Conclusions

We created two approaches for the detection and resolution of abbreviations in Dutch medical texts. Both methods, a pattern-based system and a classification-based learning system, were evaluated on a dataset of about 65,000 words which was annotated for this task [4]. For both detection tasks, the learning approach showed a performance increase over the pattern-based approach. For the detection of abbreviations, an F-score of 93% was obtained; for definition detection, the classifier obtained an F-score of 82%. In future work, we will apply decompounding techniques to refine the matching of abbreviations against word and word part boundaries. In order to tackle the problem of cross-lingual matching, viz. matching English abbreviations to Dutch expansions, we will use external sources and experiment with MT techniques. We will also try to resolve undefined abbreviations with the use of external -internet- sources and we will add the formation patterns described by Liu et al. (2001) to our pattern-based and learning system.

[4] The data sets will be made available from `http://veto.hogent.be/lt3/downloads`

6. References

D. A. Bloom. 2000. Acronyms, abbreviations and initialisms. *BJU Int*, 86(1):1–6. 1464-4096 (Print)Journal Article Review.

R. J. Byrd, Y. Ravin, and J. Prager. 1994. Lexical assistance at the information-retrieval user interface. Technical report, IBM T.J. Watson Research Center.

J. Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22:249–254.

J.T. Chang and H. Schütze. 2006. Abbreviations in Biomedical Text. In Sophia Ananiadou and John McNaught, editors, *Text Mining for Biology and Biomedicine*, pages 99–119. Artech House, Boston/London.

J.T. Chang, H. Schütze, and R. B. Altman. 2002. Creating an online dictionary of abbreviations from MEDLINE. *J Am Med Inform Assoc*, 9(6):612–620.

W. Daelemans and A. van den Bosch. 2005. *Memory-Based Language Processing*. Cambridge University Press.

C. Friedman, H. Liu, L. Shagina, S. Johnson, and G. Hripcsak. 2001. Evaluating the UMLS as a Source of Lexical Knowledge for Medical Language Processing. *Proceedings of the American Medical Informatics Association Annual Symposium*, pages 189–193.

T. Kudo and Y. Matsumoto. 2003. Fast methods for kernel-based text analysis. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 24–31.

H. Liu, Y. A. Lussier, and C. Friedman. 2001. A study of abbreviations in the UMLS. *Proc AMIA Symp*, pages 393–7. 1531-605X (Print) Evaluation Studies Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

H. Liu, A. R. Aronson, and C. Friedman. 2002. A study of abbreviations in MEDLINE abstracts. *Proc AMIA Symp*, pages 464–8. 1531-605X (Print) Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

D. Maynard and S. Ananiadou. 1999. Term Extraction using a Similarity-based Approach. In *Recent Advances in Computational Terminology*. John Benjamins.

Y. Park and R.J. Byrd. 2001. Hybrid Text Mining for Finding Abbreviations and Their Definitions.

J. Pustejovsky, J. Castaño, B. Cochran, M. Kotecki, and M. Morrell. 2001. Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo*, 10(Pt 1):371–375.

J. Pustejovsky, J. Casta no, R. Saurí, A. Rumshinsky, J. Zhang, and W. Luo. 2002. Medstract: creating large-scale information servers for biomedical libraries. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 85–92.

B. Roark and E. Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction.

K. Taghva and J. Gilbreth. 1999. Recognizing acronyms and their definitions. *IJDAR*, 1(4):191–198.

J. D. Wren and H. R. Garner. 2002. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of information in medicine*, 41(5):426–434.

H. Yu, G. Hripcsak, and C. Friedman. 2002a. Mapping abbreviations to full forms in biomedical articles. *J Am Med Inform Assoc*, 9(3):262–72. 1067-5027 (Print) Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

H. Yu, G. Hripcsak, and C. Friedman. 2002b. The Effect of Abbreviations on MEDLINE Searching. *Journal of the American Medical Informatics Association*, 9:262–272.