Towards a learning approach for abbreviation detection and resolution

Klaar Vanopstal, Bart Desmet, Véronique Hoste

LT^3, Language and Translation Technology Team
University College Ghent
{klaar.vanopstal,bart.desmet,véronique.hoste}@hogent.be

Department of Applied Mathematics & Computer Science
Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium

May 19, 2010

1 Background

1. Background

2. Annotation

1. Background
2. Annotation
3. Pattern-based approach

1. Background
2. Annotation
3. Pattern-based approach
4. Learning-based approach

1. Background
2. Annotation
3. Pattern-based approach
4. Learning-based approach
5. Conclusions and future work

Problem

Information explosion ⇒ growing number of (bio)medical abbreviations.

New abbreviations are created; not always known to the reader.

⇒ automatic detection and resolution

Use

- information retrieval
- information extraction
- NER
- anaphora resolution

Corpus

- English
  - AbbRE: reliable standard but limited size
  - Medstract: publicly available and commonly used

Corpus

- English
  - AbbRE: reliable standard but limited size
  - Medstract: publicly available and commonly used
- Dutch: no resources available

Corpus

- English
  - AbbRE: reliable standard but limited size
  - Medstract: publicly available and commonly used
- Dutch: no resources available
- Abstracts from 2 medical journals:
  - *Nederlands Tijdschrift voor Geneeskunde* (NTvG); 29,978 words
  - *Belgisch Tijdschrift voor Geneeskunde* (TvG); 36,757 words

⇒ total of 66,739 words

Different **types** of abbreviations included in annotations:

- **Truncation**

Example

adm for *administration*

Different **types** of abbreviations included in annotations:

- **Truncation**

Example

adm for *administration*

- **First letter initialization**

Example

AAA for *abdominal aortic aneurysm*

Different **types** of abbreviations included in annotations:

- **Truncation**

Example

adm for *administration*

- **First letter initialization**

Example

AAA for *abdominal aortic aneurysm*

- **Opening letter initialization**

Example

HeLa for *Henrietta Lacks*

- **Syllabic initialization**

Example

BZD for *benzodiazepine*

- **Syllabic initialization**

Example

BZD for *benzodiazepine*

- **Substitution initialization**

Example

Fe for *iron*

- **Syllabic initialization**

Example

BZD for *benzodiazepine*

- **Substitution initialization**

Example

Fe for *iron*

- **Combination of letters and numbers**

Example

CXCR4 for *chemokine receptor fusin*

Labels

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

Example

Hoge-resolutie-computertomografie (**HRCT**)
EN: High resolution computed tomography (HRCT)

Labels

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

Example

Hoge-resolutie-computertomografie (**HRCT**)
EN: High resolution computed tomography (HRCT)

2. **ABBR_DE**: Dutch abbreviations with full form in abstract (not in local context)

Example

de pathofysiologie van het **CFS**
EN: the pathophysiology of CFS

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

Example

Hoge-resolutie-computertomografie (HRCT)
EN: High resolution computed tomography (HRCT)

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

Example

Hoge-resolutie-computertomografie (HRCT)
EN: High resolution computed tomography (HRCT)

4. **ABBR_IN_COMP**: part of a compound word; no definition in the abstract

Example

HIV-patiënten
(EN: HIV patients)

5. **ABBR_IN_COMP_DE**: part of a compound word; full form in abstract

Example

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...)

5. **ABBR_IN_COMP_DE**: part of a compound word; full form in abstract

Example

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**: abbreviations without full form

Example

AIDS, HIV

7. **ABBR_EN**: English abbreviation with Dutch/English definition in local context

Example

endosonografie (**EUS**)
EN: endoscopic ultrasound (EUS)

7. **ABBR_EN**: English abbreviation with Dutch/English definition in local context

Example

endosonografie (**EUS**)
EN: endoscopic ultrasound (EUS)

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

Example

Mini Mental State Examination (**MMSE**)

⇒ Kappa score: 0.89

	NTvG	TvG
ABBR	11.60	14.25
ABBR_DE	30.62	22.55
ABBR_IN_COMP	7.14	22.43
ABBR_IN_COMP_DE	16.85	4.96
ABBR_NO_DEF	27.65	29.12
ABBR_EN	6.14	6.69
TOTAL %	3.36	2.19

Table: Labels and their frequencies in the corpus (%)

	NTvG	TvG
def: loc	17.74%	20.94 %
def: broad	47.47%	27.50%
def: loc/broad	65.21%	48.45%

Table: Abbreviations and defined abbreviations in the corpus

⇒ Between 45% and 52% of the abbreviations are undefined

Challenges

- English abbreviations with Dutch full form: no match

Example

HAART = **k**rachtige **a**ntiretrovirale **t**herapie

Challenges

- English abbreviations with Dutch full form: no match

Example

HAART = **k**rachtige **a**ntiretrovirale **t**herapie

- Parenthetical patterns

Example

gunstige uitkomst (**score 5**)

Challenges

- English abbreviations with Dutch full form: no match

Example

HAART = **k**rachtige **a**ntiretrovirale **t**herapie

- Parenthetical patterns

Example

gunstige uitkomst (**score 5**)

- Syllabic initialization

Example

CVS = **c**hronische-**v**ermoeidheids**s**yndroom
EN: CFS = **c**hronic **f**atigue **s**yndrome)

Challenges

- English abbreviations with Dutch full form: no match

Example

HAART = **k**rachtige **a**ntiretrovirale **t**herapie

- Parenthetical patterns

Example

gunstige uitkomst (**score 5**)

- Syllabic initialization

Example

CVS = **c**hronische-**v**ermoeidheids**s**yndroom
EN: CFS = **c**hronic **f**atigue **s**yndrome)

Pattern-based approach - Related research

⇒ Use of patterns to detect abbreviations:

Pattern-based approach - Related research

⇒ Use of patterns to detect abbreviations:

- short uppercase words
- typical patterns: “long form (short form)” or “short form (long form)”

Pattern-based approach - Related research

⇒ Use of patterns to detect abbreviations:

- short uppercase words
- typical patterns: "long form (short form)" or "short form (long form)"
- identification of definitions:
  - window of **2*N** (Taghva & Gilbreth, 1999) or **3*N** words (Stanford Medical Abbreviation Method (Chang & Schütze, 2006))
  - **text markers**: () " =
  - **linguistic cues**: "short", "or" (Park & Byrd, 2001)

- + use of NLP tools to refine the search space of the definitions (Pustojevski et al., 2001) and/or to tackle the problem of function word matching

Example

ADL = **a**ctiviteiten van het **d**agelijkse **l**even
EN: daily life activities

2 steps:

2 steps:

- Abbreviation detection

2 steps:

- Abbreviation detection
- Definition matching

Step 1: abbreviation detection:

Step 1: abbreviation detection:

- capital letters / combinations of capital letters with 1-3 lowercased letters or numbers

Example

QSRL
pANCA
CDG1A

Step 1: abbreviation detection:

- capital letters / combinations of capital letters with 1-3 lowercased letters or numbers

Example

QSRL
pANCA
CDG1A

- window of 3*N words

Step 1: abbreviation detection:

- capital letters / combinations of capital letters with 1-3 lowercased letters or numbers

Example

QSRL
pANCA
CDG1A

- window of 3*N words
- text markers () = “ ’ ⇒ list of candidate definitions

Step 2: definition matching:

Step 2: definition matching:

- list of candidate definitions

Step 2: definition matching:

- list of candidate definitions
- matching: first letter of abbreviation - words in candidate definition

  ⇒ matching word + rest of the 3*N sequence = definition

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: Results of the pattern-based approach

Error Analysis

- Errors in abbreviation detection step

Error Analysis

- Errors in abbreviation detection step
  - **Titles** printed in capital letters

Error Analysis

- Errors in abbreviation detection step
  - **Titles** printed in capital letters
  - **Roman numerals** confused with capitalized i, v or x

Error Analysis

- Errors in abbreviation detection step
  - **Titles** printed in capital letters
  - **Roman numerals** confused with capitalized i, v or x
  - **single letters** which are not abbreviations (e.g. hepatitis **A**)

Error Analysis

- Errors in abbreviation detection step
  - **Titles** printed in capital letters
  - **Roman numerals** confused with capitalized i, v or x
  - **single letters** which are not abbreviations (e.g. hepatitis **A**)
  - abbreviations with **word-internal capital letters** (e.g. mmHg (EN: Torr))

Error Analysis

- Errors in abbreviation detection step
  - **Titles** printed in capital letters
  - **Roman numerals** confused with capitalized i, v or x
  - **single letters** which are not abbreviations (e.g. hepatitis **A**)
  - abbreviations with **word-internal capital letters** (e.g. mmHg (EN: Torr))
  - abbreviations with **no typical orthographical characteristics** (e.g. min)

- Errors in definition matching step

- Errors in definition matching step
  - **error percolation**

- Errors in definition matching step
  - **error percolation**
  - **mislinked words** (e.g. **h**et **h**epatitis-**A**-**v**irus (HAV))

- Errors in definition matching step
  - **error percolation**
  - **mislinked words** (e.g. **h**et **h**epatitis-**A**-**v**irus (HAV))
  - **function words** (e.g. **op** evidentie gebaseerde zorg (EBZ) (EN: evidence-based medicine (EBM))

- Errors in definition matching step
  - **error percolation**
  - **mislinked words** (e.g. **h**et **h**epatitis-**A**-**v**irus (HAV))
  - **function words** (e.g. **op** evidentie gebaseerde zorg (EBZ) (EN: evidence-based medicine (EBM))
  - **English** abbreviations with a **Dutch** definition

- Errors in definition matching step
  - **error percolation**
  - **mislinked words** (e.g. **h**et **h**epatitis-**A**-**v**irus (HAV))
  - **function words** (e.g. **op** evidentie gebaseerde zorg (EBZ) (EN: evidence-based medicine (EBM))
  - **English** abbreviations with a **Dutch** definition
  - **contractions** (e.g. **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).)

Learning-based approach - Related research

- Often in combination with pattern-based techniques, e.g. Stanford Medical Abbreviation Method (2006), Chang et al. (2002)

Learning-based approach - Related research

- Often in combination with pattern-based techniques, e.g. Stanford Medical Abbreviation Method (2006), Chang et al. (2002)
- Pattern-based detection of abbreviations + learning-based matching with definitions

Learning-based approach - Related research

- Often in combination with pattern-based techniques, e.g. Stanford Medical Abbreviation Method (2006), Chang et al. (2002)
- Pattern-based detection of abbreviations + learning-based matching with definitions
- examples of features:
  - % of characters aligned at beginning of word
  - % of characters aligned on syllable boundary
  - number of words that were skipped (negative weight)

Learning-based approach - Related research

- Often in combination with pattern-based techniques, e.g. Stanford Medical Abbreviation Method (2006), Chang et al. (2002)
- Pattern-based detection of abbreviations + learning-based matching with definitions
- examples of features:
    - % of characters aligned at beginning of word
    - % of characters aligned on syllable boundary
    - number of words that were skipped (negative weight)

Own approach

- Preprocessing steps:

Own approach

- Preprocessing steps:
  - tokenization

Own approach

- Preprocessing steps:
  - tokenization
  - POS tagging + NP chunking (Daelemans & van den Bosch, 2005)

- Learning experiments

- Learning experiments
  - YamCha (Kudo & Matsumoto, 2003): open source sequence tagger using SVM

- Learning experiments
  - YamCha (Kudo & Matsumoto, 2003): open source sequence tagger using SVM
  - 10-fold cross-validation

- Feature vector:

- Feature vector:
  - token
  - POS
  - name initials
  - sentence-initial position
  - morphological features (initial capital letter, completely capitalized, internal capital letters, lowercased, roman number, punctuation, hyphens, exclusively consonants)
  - prefix and suffix information
  - symbolic word shape feature: all morphological (binary) features
  - feature to match 1st letter of abbreviation against words in 3*N sequence

Results

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

Conclusions

Conclusions

- **annotated dataset** of +/- 67,000 words (Dutch, medical)

Conclusions

- **annotated dataset** of +/- 67,000 words (Dutch, medical)
- **2 approaches**: **pattern-based** and **classification-based**

Conclusions

- **annotated dataset** of +/- 67,000 words (Dutch, medical)
- **2 approaches**: **pattern-based** and **classification-based**
- classification-based approach **outperforms** the pattern-based approach on both tasks:

Conclusions

- **annotated dataset** of +/- 67,000 words (Dutch, medical)
- **2 approaches**: **pattern-based** and **classification-based**
- classification-based approach **outperforms** the pattern-based approach on both tasks:
  - abbreviation detection: 93% F-score

Conclusions

- **annotated dataset** of +/- 67,000 words (Dutch, medical)
- **2 approaches**: **pattern-based** and **classification-based**
- classification-based approach **outperforms** the pattern-based approach on both tasks:
  - abbreviation detection: 93% F-score
  - definition matching: 82% F-score

Future work

Future work

- incorporate information from **error analysis** into learning approach

Future work

- incorporate information from **error analysis** into learning approach
- apply **decompounding** techniques (syllabic initializations)

Future work

- incorporate information from **error analysis** into learning approach
- apply **decompounding** techniques (syllabic initializations)
- **cross-lingual matching**: external sources + MT techniques

Future work

- incorporate information from **error analysis** into learning approach
- apply **decompounding** techniques (syllabic initializations)
- **cross-lingual matching**: external sources + MT techniques
- **undefined** abbreviations: external sources

Future work

- incorporate information from **error analysis** into learning approach
- apply **decompounding** techniques (syllabic initializations)
- **cross-lingual matching**: external sources + MT techniques
- **undefined** abbreviations: external sources
- **F-scores** per label (now focus on abbreviations and definitions)

Future work

- incorporate information from **error analysis** into learning approach
- apply **decompounding** techniques (syllabic initializations)
- **cross-lingual matching**: external sources + MT techniques
- **undefined** abbreviations: external sources
- **F-scores** per label (now focus on abbreviations and definitions)
- **English** corpus