Parsing Early Modern English corpora

Automatic annotation and normalisation

Gerold Schneider, Hans Martin Lehmann, Peter Schneider
Adapted version of ICEHL 2012 talk, Zürich

1. INTRODUCTION

Data

<table>
<thead>
<tr>
<th>source</th>
<th>word-count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archer 3.2</td>
<td>~3,200,000</td>
</tr>
<tr>
<td>ZEN 1.0</td>
<td>1,796,989</td>
</tr>
<tr>
<td>Old Bailey Corpus</td>
<td>~14,000,000</td>
</tr>
<tr>
<td>total</td>
<td>~19,000,000</td>
</tr>
</tbody>
</table>

Archer Corpus
http://www.llc.manchester.ac.uk/research/projects/archer/

ZEN Corpus
http://www.es.uzh.ch/zen

Old Bailey Corpus
http://www.uni-giessen.de/oldbaileycorpus/index.html

Annotation: Tagging, Lemmatization, Chunking

Annotation: Parsing

Pro3Gres parser
Hand-written competence grammar
Penn Treebank performance disambiguation
Set of dependency labels similar and mappable to Stanford scheme
**Parsing: Tri/Bi-lexicalization**

The parser estimates the probability of the dependency relation $R$ at distance (in chunks) $\text{dist}(a, b)$, given the lexical head $a$ of the governor and the lexical head $b$ of the dependent.

\[
p(R, \text{dist}(a, b)) = P(R|a, b) \cdot P(\text{dist}|R, a, b)
\]

- $\text{eat pizza} \ vs. \ \text{eat yesterday}$; $\text{take report into account} \ vs. \ \text{take report on account}$
- Backed off via WordNet (top level) and mono-lexicalized model
- Uses beam search, non-greedy
- Context-free CKY parser (CKY = shift & reduce, tabular approach)
- Many statistical extensions
- MLE models
- Adding distributional semantics, self-training, etc.

2. **NORMALISATION**

Frequent problems for automatic annotation:

- $\_d \to -ed$,
- $\_t \to -ed$,
- $\_es \to -s$.

C&C-tagger

---

**Parsing: Evaluation**

**General Evaluation**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Subject</th>
<th>Object</th>
<th>Noun-PP</th>
<th>Verb-PP</th>
<th>sub.clause</th>
</tr>
</thead>
</table>
| **GREval** | Precision | 92% | 89% | 74% | 72% | 74%
| Recall | 81% | 84% | 65% | 85% | 62% |
| **GENIA** | Precision | 90% | 93% | 85% | 82% |
| Recall | 91% | 91% | 82% | 84% |
| **BNC-w old** | Precision | 86% | 87% | 89% |
| Recall | 83% | 88% | 70% |
| **BNC-x new** | Precision | 89% | 75% | 75% | 83% | 73% |
| Recall | 86% | 83% | 77% | 69% | 63% |

- **GREval**: 500 random sentences from Suzanne corpus
- **GENIA**: 100 random sentences from biomedical GENIA corpus
- **BNC (written part)**: 100 random sentences annotated by us
  - old: Treetagger-Carafe Pipeline / BNC World
  - new: LT-TTT2 Pipeline / BNC XML (different rand set used for old & new)
Applying VARD to the ZEN corpus

Test run using default setup folder (auto-normalise 50% threshold)

Manual cursory analysis of VARD output
- Most error concern place names and proper names
  - St. Paul’s cathedral -> St. Pal’s cathedral

Improve output via preprocessing and training
- Using VARD text-to-ignore to construct regular expressions to skip probable names
- Skip ZEN text rendition tags (italics)
  Manual identifying and correcting most frequent types
  - Check suggested variant types
  - Check non-identified types
Re-run VARD auto-normalise 50% threshold

Context sensitive pre-processing (text-to-ignore)

skip text in foreign languages: <foreign[^<]*>(.)+?<\foreign>
skip some place names:
\b([Cc]ounty|[Cc]ourt|States?|Country|Town|Borough|Province|Parish|Vicarage|Island|Castle|Garrison|late)\s+of\s+(St\.| the)\s+(A-Z)[a-z]+\s+\s+\b
skip some common/place names (title sequence, of/de/d'):
\b\((\s+Count(ess)?|Earl|Prince|Lord|Mare\s?|Mrs\s?|\s+Sir|Monsieur|Princess|Dukes?|Gouvernou\s?|Dean|Cardinal|Doctor|Justice|Principe|Sei\s+|Chevalier|Marquis|Baron)\s+\s+\s+(of|de\s+|the)\s+(A-Z)[a-z]+\s+\s+\b
skip some place names with mandatory “of”:
\b(King|Queen|Bishop|Archbishop)\s+of\s+(A-Z)[a-z]+\s+\b
skip some personal names with postponed titles:
\b\((A-Z)[a-z]+\s+\s+\s+(Kt|Esq|Esquire|Barone)?\s+\s+\b
Skip names after abbreviated Saint: \b\((St|ST)\s+\s+\s+\s+(A-Z)[a-z]+\s+\s+\s+\s+\b
skip graphical entities: (&[lmn]dash;)+ (&[Dd]agger;)+ (&[Hh]and;)+
Specific vocabulary training

Standard word list is based on most frequent items in modern corpora
Missing dictionary items lead to over-eager normalisation (precision errors)

Assignee -> assigns
Patience -> Patience
Relict -> relic
Footpad -> Footpath
Porte -> Port (central government of the Ottoman Empire)
Dom -> Doom (Spanish title, e.g. Dom Fernando de Prado)
Medical (advertising): ureter, gleet
Professions/crafts: Hosier, Mercer

Disused forms: thou, thee, thine; hath, doth, taketh (recall errors)
– Should VARD replace with modern equivalent?

Usability Constraints, pitfalls

Interactive mode provides good overview but gets slow with large files
– Edit variant replacement and word list manually
– Rules.txt limited usability: Handling of hyphen
  – Cannot set up a rule for automatically splitting hyphenated items into two words
    – E.g. Name-street|lane|row|fields -> Name street|lane|row|fields
  – Hyphen -> two word rule could enable further normalisation of each w-unit:
    – Contribution-mony -> contribution money

To do list

Abbreviations
– Titles: The Rt. Hon. XY (Right Honourable XY)
– First names: Geo. (George) Wm. (William) Tho. (Thomas)
– Shorthand: &c.
– Double use of full stop as s-unit marker and abbreviation marker
– Full stop needs to be removed unless at end of s-unit
– Abbreviations skipped via text-to-ignore need to be expanded by post-processing
  – Capt. -> Captain; Lieut. -> Lieutenant;
Plural s’ genitive: ‘ should not be removed
– Ladies’ -> Ladies
Oops

leaper x> leper (leaper)
A BAY GELDING, got by Tandem, six years old, 15 hands and an inch, a
remarkable good <normalised orig="leaper" auto="true">leper</normalised>, a
very fleet good hunter, and master of 15 stanes to any hounds;
s genitives with apostrophe of words ending in –s
... and cut one <normalised orig="another's" auto="true">another</normalised> throats for nothing
Sould x> sold (should)
<normalised orig="Sould" auto="true">Sold</normalised> hanging continue
fashionable, it will not operate as a terror to the multitude.
Noun meaning
He called his friend, the <normalised orig="Abbe" auto="true">Abbey</normalised> Taleyrand (Bishop of Autun)
changed vowel from u -> i instead of metathesis ur -> ru
whereby the Devil does <normalised orig="thurst" auto="true">thirst</normalised> them either into Desperation, or into Wretchedness

Tagging Accuracy / Outlook

Tagging Accuracy

Rayson et al. (2007) report an increase of about 3% (from 82% to 85%
accuracy) on tagging Shakespeare texts.
As an upper bound, when tests are manually normalized, they report 89%
accuracy.

Outlook

Spelling Normalisation can be seen as a translation task between closely
related related languages. Could character-level SMT (Tiedemann, this workshop)
be applied?

3. PARSING

3.1 First impressions
• Early modern English has long and complex sentences

3.2 How much does normalisation improve parsing?
• From non-normalised to VARDed standard
• From VARDed standard to VARDed re-trained

3.3 Parser Adapation
• Error Analysis
• Closed-class lexis extensions
• More semantics and context
• A word on distributional semantics and self-training
• More collocations, freer word order
Parsing – First Impressions

First glance at texts from 16xx and 17xx and their automatic parses suggests

• more conjunction problems: is safely arrived\_VBN and prosecuting\_VBG (past and present participle)
• sentences are extremely long / sentences do not exist as a concept
• semicolons often have sentence-boundary function \(\rightarrow\) break longer sentences at each semi-colon
• freer constituent order
• as consequence, even more ambiguity / more semantic knowledge needed
• ellipsis / copula-less constructions: She not very well. Impossible to analyse in current dependency format \(\rightarrow\) not annotated

Example sentence (Archer 16xx, VARDed)

\begin{quote}
butt she was well built, a fair ship, of a good burden, and had mounted in her forty pieces of brass cannon, two of them demi cannon, and she was well manned, and of good force and strength for war: she was a good sailor, and would turn and tack about well; she held 100 persons of Whitelocke’s followers, and most of his baggage, besides her own mariners, about 200.
\end{quote}

Genitive of quality / quality of PP: of a good burden, of good force
X-bar violation: mounted \[in her\] forty pieces
Ellipsis: two of them demi cannon
Conjunctions: was well manned, and of good force and strength for war (adjective/participle and complex PP in coordination)
Appositions: besides her own mariners, about 200
Generally, even more ambiguity \(\rightarrow\) similar to PDE but harder (more frequent, "statistical difference").

3.1 How much does normalisation improve parsing?

From unnormalised to VARDed standard

In a 100 sentences random sample from ARCHER 16xx: 131 normalisations are made (batch, 50% confidence).
• Better syntactic analysis due to VARD: 12 sentences
• Worse syntactic analysis due to VARD: 1 sentence
• Corrections, but also new errors: 3 sentences

In a random 422 sentence sample from ZEN: 332 sentences contain differences due to VARDing.

<table>
<thead>
<tr>
<th></th>
<th>Better</th>
<th>Worse</th>
<th>Equal</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>21</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>obj</td>
<td>17</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>pobj</td>
<td>10</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>modpp</td>
<td>9</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>sentobj</td>
<td>11</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>sum</td>
<td>68</td>
<td>5</td>
<td>139</td>
</tr>
</tbody>
</table>
3.1 How much does normalisation improve parsing?

From VARDed standard to VARDed retrained

In the random 422 sentence sample from ZEN: 132 sentences contain differences due to our retraining VARD.

<table>
<thead>
<tr>
<th></th>
<th>Better</th>
<th>Worse</th>
<th>Equal</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>4</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>obj</td>
<td>2</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>pobj</td>
<td>2</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>modpp</td>
<td>3</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>sentobj</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Σ</td>
<td>11</td>
<td>1</td>
<td>45</td>
</tr>
</tbody>
</table>

Example

With VARD standard Normalisation

With VARD re-trained Normalisation

3.2 Parser Adaptation – Error Analysis

Based on the manual annotation of 25 sentences from each century

From 16xx:
1. pobj(mount, her, in, 1). x-bar violation: already chunker fails

But #56: subj=through#44=comp->he#45=subject->

But CL though_IN he_PRP

formerly_RB done_VBN

good_3
the_DT

Services_NNS for_IN City_NNP

But #46: subj=have#46

But CL though_IN he_PRP

formerly_RB done_VBN

good_3
the_DT

Services_NNS for_IN City_NNP

Example

With VARD standard Normalisation

With VARD re-trained Normalisation

3.2 Parser Adaptation – Error Analysis

2. modpp(ship, burden, of, 1).

but she was well built, a fair ship, of a good burden

unusual/unseen noun-PP combination = quality of-PP: commas are rare → triggers apposition reading

3. subj(reinforce, state, 12).

unusual/unseen adverbs inside the verb group: chunker fails [NB: late standardisation]
3.2 Parser Adaptation – Error Analysis

4. **pobj**(be,cabin,in,2).
5. **pobj**(be,make,of,2).

The cabins wherein _NN Whitelocke was, were of an handsome _VB make. 

Wherein is rare/unknown as relative pronoun: gets tagged as noun
Grammar does not know this relative pronoun and would not get the LDD
(anaphor resolution) to cabin
Make is rare/unknown as noun: gets tagged as verb

6. **sentobj**

that, for his part, he had done nothing but sigh for her ever since she came; and
that all the white beauties he had seen, never charmed him so absolutely as this
fine creature had done; and that no man of any nation, ever beheld her, that did
not fall in love with her; and that she had all the slaves perpetually at her feet;
and the whole Country resounded with the fame of Clemene; “For so,” said he,
“we have Christened her.”

It is not clear to the annotator which sentences are subordinate and which
are coordinate. Related: concept of sentence was less standardised.

---

29.05.13  Page 29

---

3.2 Parser Adaptation – Error Analysis

From 17xx:

1. **obj**(have,little,16).

He is such an itinerant, to speak that I have but little of his company.

But is not known as adverb: gets tagged as conjunction

---

Closed-class lexis extensions

Some closed-class words, e.g. _but_ as adverb, and _lest_, are not known to the
grammar

—but is not known as adverb: gets tagged as conjunction

— add dedicated underchunking rule

These adaptations are simple and work well, but only lead to
small, very specific improvements.
More semantics and context

The ambiguity trade-off: constrain <-> disambiguate

- Constrain too much: correct reading cannot be found
- Constrain too little: ambiguity explodes, risk for incorrect disambiguation

One way to help disambiguation is to include more semantics and context:

1. Semantic expectation

Original parser models probabilities using only those syntactic relations that are in competition. E.g. objects (e.g. *eat pizza*) and nominal adjuncts (e.g. *eat Friday*) are modeled as being in competition, but not subjects and objects. Original parser models syntactic competition. We now add semantic competition: every relation is in competition with every other relation.

A sentence like the rabbit chased the dog now gets a lower probability than the dog chased the rabbit -- rabbits are very unlikely to be subjects of active instances of chase.

Our semantic world knowledge (e.g. selectional restrictions) becomes part of the model.

More semantics and context

2. Broader context

PP-attachment is typically most ambiguous. The interaction between different PPs was not considered in the original statistical model. World knowledge expressed across more than one node generation was lost. Now: probability that PP2 is a dependent of PP1 (PP1 < PP2) in a verb-PP-PP sequence, given the lexical items, is calculated as follows:

\[
p(\text{verb} < (PP_1 < PP_2)) = \frac{\#(\text{verb} < (PP_1 < PP_2))}{\#(\text{verb} < (PP_1 < PP_2)) + \#(\text{verb} < PP_1 < PP_2)}
\]
A word on distributional semantics

Non-negative matrix factorization (Lee and Seung, 2001) is similar to LSA. It also boosts plausible but unseen combinations.

- Used in biological phylogeny, e.g. (Murrell et al., 2011)
- Never uses negative weights → suitable for treating probabilities

For our verb+noun PP attachment probability matrix

- Initial verb-prep and noun-prep matrices are filled with attachment probabilities. Null counts are given p=0.2
- A version for multiple PPs (verb-prep1-prep2 ; noun-prep1-prep2) is also produced (probability of prep2 attaching to noun or verb)
- The approximated matrix contains non-sparse (non-zero) probabilities for every verb and every preposition.

<table>
<thead>
<tr>
<th>PP-attachment</th>
<th>with MLE multi-PP = Base system</th>
<th>with NNMF multi-PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>nounpp_prec</td>
<td>354 of 492 71.95%</td>
<td>358 of 499 71.73%</td>
</tr>
<tr>
<td>verbpp_prec</td>
<td>357 of 481 74.22%</td>
<td>355 of 471 75.37%</td>
</tr>
<tr>
<td>ncmod.recall</td>
<td>536 of 801 66.92%</td>
<td>538 of 801 67.17%</td>
</tr>
<tr>
<td>iobj.recall</td>
<td>140 of 157 89.17%</td>
<td>140 of 157 89.17%</td>
</tr>
<tr>
<td>argmod.recall</td>
<td>35 of 40 87.5%</td>
<td>34 of 40 85.0%</td>
</tr>
</tbody>
</table>

Above Extensions & BNC Self-Training on PDE

Pro3Gres has a very strong correlation between backoff level and parser accuracy. Fully lexicalized decisions (Level 0: head + preposition + description noun), have much higher performance than those further down the back-off chain. Level 2 is verb + preposition, level 3 is head class + preposition + noun, level 4 is head class + preposition + description noun class, level 5 is preposition + description noun class, level 6 is preposition only.

We test if there are semantics-based methods to reduce sparseness → more decisions can be taken at early backoff levels.

Dependency Bank

take -> obj -> decision

Outlook: Let's use the parsed Google-Books corpus for self-training. Thx to Joakim
More collocations, freer word order

Freer word order necessitates removing constraints
Thus leads to many wrong analyses, which requires better ranking
Already the original statistics model has implicit collocation measure
\[ p(\text{REL} | \text{governor word}, \text{dependent word}) \]
We have added an explicit, simple collocation measure
\[ \text{Delta } P = p(w1|w2) - p(w1) \] (see e.g. Gries 2012)
Particularly for content words, \( p(w1) \) is small
We use \( p(\text{governor word} | \text{dependent word}) \)
The resulting stats model is a bit more robust to removing constraints
PPs to the left are allowed more generally
The resulting parser has the best overall recall (not best total performance)

Conclusions

- Parsing quality is and stays lower: ~ 1/3 more errors on 16xx than 19xx
- Spelling normalisation helps most
- VARD is a useful tool, some pre-processing / suggestions
- Lower tagging error rate leads to fewer parsing errors (expected)
- Closed class extensions to the grammar are simple but local
- Ambiguity is even higher in earlier texts
- Otherwise, few grammatically really different constructions (“statistical difference”)
- Adding more context and semantics helps to reduce errors
- New perspectives for diachronic descriptive linguistics:
  - Measure signal
  - Inspect parsing errors

Questions

7 AC2:1414 How did he handle the questions?
8 AHY:97 The server does all the work of finding the information so in theory the user can
9 AKH:751 And how do we answer this question?
10 AMB:559 Why do you ask so many questions, you irritating little brat?
11 ANH:621 or should one ask a separate question regarding its authority in matters of tax concerning its authority in matters of corporate taxation and personal taxation?
12 APP:496 Had the Tuairisc model been adopted there would have been no question of Niall Derry: this would have been the responsibility of purely local groups.
13 AYK:1794 How often do we ask the question Why me?
14 AYM:1268 In the meantime should you have any questions, please ring The National Dire
Verb normalizations per decade in ZEN

(vnorm per 10'000 words)