Automatic Abstract Anaphora Resolution in German

Adam Roussel
September 13, 2017
Ruhr-Universität Bochum
Introduction
Some examples

What do I mean by ‘abstract anaphora’\(^1\)?

(1) Hammond wanted to bring some dinosaurs back to life. Dr. Grant thought this was a bad idea.

\(^1\)Aka. discourse deixis (Webber, 1988), complex anaphors (Consten, Knees, & Schwarz-Friesel, 2007) ...
What do I mean by ‘abstract anaphora’? \(^1\)

(1) Hammond wanted to bring some dinosaurs back to life. Dr. Grant thought this was a bad idea.

(2) But Hammond was determined to do it.

\(^{1}\) Aka. discourse deixis (Webber, 1988), complex anaphors (Consten et al., 2007) …
What do I mean by ‘abstract anaphora’?\(^1\)

(1) Hammond wanted to bring some dinosaurs back to life. Dr. Grant thought *this* was a bad idea.

(2) But Hammond was determined to do *it*.

(3) Dr. Grant was not aware of the *fact* that he had already done it.

---

\(^1\)Aka. discourse deixis (Webber, 1988), complex anaphors (Consten et al., 2007) ...
Some examples

What do I mean by ‘abstract anaphora’?\(^{1}\)

(1) Hammond wanted to bring some dinosaurs back to life. Dr. Grant thought \textit{this} was a bad idea.
(2) But Hammond was determined to do \textit{it}.
(3) Dr. Grant was not aware of the \textit{fact} that he had already done it.
(4) Dr. Sattler was not aware of \textit{this} / \textit{this fact} either.

\[^{1}\text{Aka. discourse deixis (Webber, 1988), complex anaphors (Consten et al., 2007) …}\]
Some examples

What do I mean by ‘abstract anaphora’\(^1\)?

(1) Hammond wanted to bring some dinosaurs back to life. Dr. Grant thought this was a bad idea.
(2) But Hammond was determined to do it.
(3) Dr. Grant was not aware of the fact that he had already done it.
(4) Dr. Sattler was not aware of this / this fact either.

\(^1\)Aka. discourse deixis (Webber, 1988), complex anaphors (Consten et al., 2007) ...
Some examples

What do I mean by ‘abstract anaphora’\(^1\)?

(1) Hammond wanted to bring some dinosaurs back to life. Dr. Grant thought \textit{this} was a bad idea.

(2) But Hammond was determined to do \textit{it}.

(3) Dr. Grant was not aware of the \textit{fact} that he had already done it. → \textit{cataphoric}

(4) Dr. Sattler was not aware of \textit{this} / \textit{this fact} either. → \textit{anaphoric}

\(^1\)Aka. discourse deixis (Webber, 1988), complex anaphors (Consten et al., 2007) ...
This system covers two broad classes of German abstract anaphors:

• Pronouns: dies ‘this’, das ‘that’, and es ‘it’
• Shell nouns: Tatsache ‘fact’, Frage ‘question’, Problem ‘problem’, etc.

Both as anaphora and cataphora

Antecedents (or catacedents) may have either verbal or nominal heads
System Design
Classification and Resolution

Usual approach involves two steps:

1. **Classification**: Deciding whether or not an instance requires resolution
2. **Resolution**: Deciding which potential antecedent belongs with a given anaphor instance
Motivation

- Need to reduce number and variety of candidates
- Effectiveness of sieve-based approaches in coreference resolution (Lee et al., 2013)
- Tendencies of shell nouns to prefer certain patterns (Schmid, 2000)
- Indications that annotator’s behavior can be approximated by relatively simple heuristics (Artstein & Poesio, 2006)
die Frage, ob das so bleiben muss

(‘the question whether it has to stay that way’)

Procedure
### Extraction Patterns

<table>
<thead>
<tr>
<th>Name</th>
<th>Schema</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN-ist-dass</td>
<td>sein → NN₁</td>
<td>Tatsache₁ ist, dass das nicht funktioniert₂</td>
</tr>
<tr>
<td></td>
<td>sein → x₂</td>
<td></td>
</tr>
<tr>
<td></td>
<td>x₂ → dass</td>
<td></td>
</tr>
<tr>
<td>NN-KOUS</td>
<td>NN₁ → V₂</td>
<td>die Frage₁, ob das so bleiben muss₂</td>
</tr>
<tr>
<td></td>
<td>V₂ → KOU</td>
<td></td>
</tr>
<tr>
<td>PDS-last-verb</td>
<td>V₂ ... {das</td>
<td>dies}₁</td>
</tr>
<tr>
<td>PDAT-last-sent</td>
<td>ROOT₂ $. NN₁</td>
<td>Er hat₂ es schon getan. Diese Tatsache₁ war ihr nicht be-wusst.</td>
</tr>
<tr>
<td></td>
<td>NN₁ → PDAT</td>
<td></td>
</tr>
</tbody>
</table>

- This implementation includes 14 such extraction patterns
- Patterns ordered according to accuracy/specificity
NN-KOUS

die Frage, ob das so muss ($, KOUS, PDS, ADV, VVINF, VMFIN)

(‘the question whether it has to stay that way’)
Prozedur

NN-KOUS

die, ob das so bleiben muss
⇒ extracted anaphor–antecedent pair: (Frage, muss)
Classification Features

- Lemmas
- Germanet features, e.g.:
  - Semantic field of anaphor and mother of anaphor
  - Whether verbal mother of anaphor could also take a clausal complement
- Syntactic features, e.g.:
  - Distance between anaphor & antecedent
  - Grammatical relation of anaphor & antecedent
  - Whether anaphor and antecedent have the same grammatical relation
  - Type of determiner of anaphor, if present
- Surface features, e.g.:
  - Whether head ends in -ung, -keit, or -heit (nominal antecedents)
  - Whether head ends in -en (substantivized verbs)
Parsed sentence

NN-dass
Architecture

Parsed sentence

NN-dass

yes

no

NN-ist-dass

yes

no

yes

NN-dass-Classifier

yes

accept

NN-ist-dass-Classifier
Evaluation
Stems from two annotation projects:

- Dipper and Zinsmeister (2012) annotating the pronouns *dies*, *das*, and *es*
- Simonjetz and Roussel (2016) annotating German (and English) shell nouns
Stems from two annotation projects:

- Dipper and Zinsmeister (2012) annotating the pronouns *dies*, *
das*, and *es*
- Simonjetz and Roussel (2016) annotating German (and English) shell nouns

Corpus used here contains 1734 annotated German instances of abstract reference. Of these,

- 1086 are shell nouns
- 249 involve either *dies* or *das*, and
- 375 involve instances of *es*.
- The remaining 24 instances involve pronominal adverbs, such as *deshalb*, and are not covered by this study.
• Comparisons with existing work
  • Classification → Was the instance correctly approved by some classifier/pattern?
  • Resolution → Was the instance correctly approved by the correct classifier/pattern?

• $F_1$ scores for classification

• Only accuracy for resolution
Classifiers and Baselines

![Graph showing precision and recall for classifiers and baselines, with a point labeled 'Constant (F_1 = 0.204)'](image)

- **Axes:**
  - **Horizontal Axis:** Recall (values range from 0.1 to 0.7)
  - **Vertical Axis:** Precision (values range from 0.1 to 0.8)

- **Points:**
  - A point labeled 'Constant (F_1 = 0.204)' indicates the performance of a baseline classifier.
Classifiers and Baselines

![Classification Performance Chart]

- **Random** (F₁ = 0.225)
- **Constant** (F₁ = 0.204)
Classifiers and Baselines

![Graph showing precision vs recall for different classifiers.]

- Stratified ($F_1 = 0.309$)
- Random ($F_1 = 0.225$)
- Constant ($F_1 = 0.204$)
Classifiers and Baselines

- SVC, C=10 ($F_1 = 0.238$)
- Stratified ($F_1 = 0.309$)
- Random ($F_1 = 0.225$)
- Constant ($F_1 = 0.204$)
Classifiers and Baselines

- **SVC, C=10** ($F_1 = 0.238$)
- **MultinomialNB, α=0.1** ($F_1 = 0.393$)
- **Stratified** ($F_1 = 0.309$)
- **Random** ($F_1 = 0.225$)
- **Constant** ($F_1 = 0.204$)
Classifiers and Baselines

- **SVC, C=10 (F₁ = 0.238)**
- **MultinomialNLP, q=0.1 (F₁ = 0.393)**
- **Logistic Regression, C=10 (F₁ = 0.304)**
- **Stratified (F₁ = 0.309)**
- **Random (F₁ = 0.225)**
- **Constant (F₁ = 0.204)**
• Most similar study, Kolhatkar and Hirst (2014):
  • Baseline (patterns alone), 57%
  • Additional heuristics, 69%

• This system:
  • Baseline (patterns alone): $F_1 = 0.104$, $P = 0.059$, $R = 0.478$
  • Baseline resolution accuracy, 72.2%
  • $\Rightarrow$ 34.5% overall
  • Best classifier: $F_1 = 0.413$, $P = 0.899$, $R = 0.272$
  • With resolution accuracy of 87%
  • $\Rightarrow$ 23.7% overall
Anaphoric Shell Nouns

• Most similar study, Kolhatkar and Hirst (2012) examined instances of this issue
  • Baseline (adjacent sentence): 22.93%
  • With classifier, 59.91%

• This system:
  • Baseline classifier: $F_1 = 0.041$, $P = 0.021$, $R = 0.640$
  • Baseline resolution accuracy, 14.00%
  • $\Rightarrow$ 8.96% overall
  • Best classifier, $F_1 = 0.263$, $P = 0.354$, $R = 0.214$
  • Resolution accuracy of 30.7%
  • $\Rightarrow$ 6.57% overall
• Most similar study, Jauhar et al. (2015):
  • Classification Baseline: $F_1 = 0.217$, $P = 0.121$, $R = 1.000$
  • Class. + Resolution Baseline: $F_1 = 0.165$, $P = 0.153$, $R = 0.179$
  • Classification: $F_1 = 0.386$, $P = 0.352$, $R = 0.429$
  • Class. + Resolution: $F_1 = 0.222$, $P = 0.226$, $R = 0.218$

• This system:
  • Baseline: $F_1 = 0.590$, $P = 0.430$, $R = 0.946$
  • Baseline resolution accuracy: 12.6%
  • Best classifier, $F_1 = 0.762$, $P = 0.691$, $R = 0.853$
  • Resolution accuracy of 15.4%
  • $\Rightarrow$ 13.14% overall
Extraction pattern errors (Pronouns)
Extraction pattern errors (Shell nouns)
• Overall, 20–25% of cases correctly assigned some antecedent; of these cases, 50–55% contain the correct antecedent
Overall, 20–25% of cases correctly assigned some antecedent; of these cases, 50–55% contain the correct antecedent.

System shows least improvement over baseline for pronouns.

Lack of relevant features.

These features work much better for shell nouns.

Ideas for future work:

- Better features targeting pronouns
- Mixing data between similar patterns (e.g., this NN)
- Integrating NP coreference information
Summary

• Overall, 20–25% of cases correctly assigned some antecedent; of these cases, 50–55% contain the correct antecedent

• System shows least improvement over baseline for pronouns
  → Lack of relevant features

• These features work much better for shell nouns

• Ideas for future work:
  • Better features targeting pronouns
  • Mixing data between similar patterns (i.e., this ≈ this NN)
  • Integrating NP coreference information
Thanks!

https://github.com/ajroussel/aaarg


Feature Set Comparison

![Graph showing precision-recall plot with data points for different feature sets. The points are labeled as follows: Only Lemmas (F₁ = 0.136), No GermaNet (F₁ = 0.318), Lemmas, etc. (F₁ = 0.309), All (F₁ = 0.363), and No lemmas (F₁ = 0.297).]
## Complete Feature Set

<table>
<thead>
<tr>
<th>Feature</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anaphor</strong></td>
<td></td>
</tr>
<tr>
<td>Lemma</td>
<td><em>das, es, Umstand</em></td>
</tr>
<tr>
<td>Number</td>
<td><em>Sing./Pl.</em></td>
</tr>
<tr>
<td>Grammatical function</td>
<td><em>subj, obja</em></td>
</tr>
<tr>
<td>Whether parent precedes anaphor</td>
<td><em>Yes/No</em></td>
</tr>
<tr>
<td>Whether parent is subjunctive</td>
<td><em>Yes/No</em></td>
</tr>
<tr>
<td>Whether parent is clausal verb</td>
<td><em>Yes/No</em></td>
</tr>
<tr>
<td>Semantic field</td>
<td><em>Attribut, Kommunikation</em></td>
</tr>
<tr>
<td>Parent semantic field</td>
<td><em>Gefühl, Perzeption</em></td>
</tr>
<tr>
<td>Semantic fields of dep. adjectives</td>
<td><em>Bewegung, Menge</em></td>
</tr>
<tr>
<td>Whether dep. article is definite or indefinite</td>
<td><em>Yes/No</em></td>
</tr>
<tr>
<td>Dep. determiners</td>
<td><em>dieser, kein, beiden</em></td>
</tr>
</tbody>
</table>
## Complete Feature Set

### Antecedent

<table>
<thead>
<tr>
<th>Feature</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent preposition lemmas</td>
<td>zu, für, nach</td>
</tr>
<tr>
<td>Dependent complementizers</td>
<td>dass, ob, weil</td>
</tr>
<tr>
<td>Grammatical function</td>
<td>root, objc</td>
</tr>
<tr>
<td>Length</td>
<td>No. of tokens</td>
</tr>
<tr>
<td>Gender</td>
<td>Masc, Fem, Neut</td>
</tr>
<tr>
<td>Semantic field</td>
<td>Attribut, Kommunikation</td>
</tr>
<tr>
<td>Embedding depth</td>
<td>No. of deps. to sentence root</td>
</tr>
<tr>
<td>If nominal, ending</td>
<td>-ung, -heit, -en</td>
</tr>
<tr>
<td>Whether antecedent contains question mark</td>
<td>Yes/No</td>
</tr>
</tbody>
</table>

### Relation

<table>
<thead>
<tr>
<th>Feature</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance between anaphor/antecedent</td>
<td>No. of tokens</td>
</tr>
<tr>
<td>Whether anaphor precedes antecedent</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Whether anaphor/antecedent funcs. match</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Whether colon between anaphor/antecedent</td>
<td>Yes/No</td>
</tr>
</tbody>
</table>
Abstract Anaphora Distribution
### Classifiers and Baselines

<table>
<thead>
<tr>
<th>Name</th>
<th>Classification</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Constant</td>
<td>0.121</td>
<td>0.654</td>
</tr>
<tr>
<td>Random</td>
<td>0.155</td>
<td>0.446</td>
</tr>
<tr>
<td>Stratified</td>
<td>0.518</td>
<td>0.234</td>
</tr>
<tr>
<td>MultinomialNB, $\alpha = 0.1$</td>
<td>0.674</td>
<td>0.244</td>
</tr>
<tr>
<td>SVC, $C = 10$</td>
<td>0.742</td>
<td>0.132</td>
</tr>
<tr>
<td>Logistic Regression, $C = 10$</td>
<td>0.722</td>
<td>0.185</td>
</tr>
<tr>
<td>Voting</td>
<td>0.774</td>
<td>0.179</td>
</tr>
</tbody>
</table>

**Table 1:** Classification performance and resolution accuracy
### Table 2: System performance for cataphoric shell noun instances

<table>
<thead>
<tr>
<th>Name</th>
<th>Classification</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Constant</td>
<td>0.059</td>
<td><strong>0.478</strong></td>
</tr>
<tr>
<td>Random</td>
<td>0.062</td>
<td>0.255</td>
</tr>
<tr>
<td>Stratified</td>
<td>0.233</td>
<td>0.102</td>
</tr>
<tr>
<td>MultinomialNB, $\alpha = 0.1$</td>
<td>0.729</td>
<td>0.269</td>
</tr>
<tr>
<td>SVC, $C = 10$</td>
<td>0.736</td>
<td>0.128</td>
</tr>
<tr>
<td>Logistic Regression, $C = 10$</td>
<td><strong>0.899</strong></td>
<td>0.272</td>
</tr>
<tr>
<td>Voting</td>
<td>0.893</td>
<td>0.254</td>
</tr>
</tbody>
</table>
## Anaphoric Shell Nouns

<table>
<thead>
<tr>
<th>Name</th>
<th>Classification</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Constant</td>
<td>0.021</td>
<td><strong>0.640</strong></td>
</tr>
<tr>
<td>Random</td>
<td>0.020</td>
<td>0.272</td>
</tr>
<tr>
<td>Stratified</td>
<td>0.056</td>
<td>0.083</td>
</tr>
<tr>
<td>MultinomialNB, $\alpha = 0.1$</td>
<td>0.354</td>
<td>0.214</td>
</tr>
<tr>
<td>SVC, $C = 10$</td>
<td>0.000</td>
<td>NaN</td>
</tr>
<tr>
<td>Logistic Regression, $C = 10$</td>
<td><strong>0.442</strong></td>
<td>0.119</td>
</tr>
<tr>
<td>Voting</td>
<td>0.428</td>
<td>0.109</td>
</tr>
</tbody>
</table>

### Table 3: System performance for anaphoric shell noun instances
## Pronouns

<table>
<thead>
<tr>
<th>Name</th>
<th>Classification</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Constant</td>
<td>0.430</td>
<td>0.946</td>
</tr>
<tr>
<td>Random</td>
<td>0.499</td>
<td>0.767</td>
</tr>
<tr>
<td>Stratified</td>
<td>0.658</td>
<td>0.835</td>
</tr>
<tr>
<td>MultinomialNB, $\alpha = 0.1$</td>
<td>0.691</td>
<td>0.819</td>
</tr>
<tr>
<td>SVC, $C = 10$</td>
<td>0.691</td>
<td>0.853</td>
</tr>
<tr>
<td>Logistic Regression, $C = 10$</td>
<td>0.693</td>
<td>0.808</td>
</tr>
<tr>
<td>Voting</td>
<td>0.697</td>
<td>0.820</td>
</tr>
</tbody>
</table>

**Table 4:** System performance for pronominal abstract anaphora
## Per-anaphor classification performance

<table>
<thead>
<tr>
<th>Anaphor</th>
<th>N</th>
<th>Recall</th>
<th>Precision</th>
<th>(F_1)-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zusicherung ‘pledge’</td>
<td>1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Notwendigkeit ‘need’</td>
<td>7</td>
<td>0.714</td>
<td>1.000</td>
<td>0.833</td>
</tr>
<tr>
<td>Tatsache ‘fact’</td>
<td>21</td>
<td>0.684</td>
<td>1.000</td>
<td>0.813</td>
</tr>
<tr>
<td>Überzeugung ‘conviction’</td>
<td>8</td>
<td>0.667</td>
<td>1.000</td>
<td>0.800</td>
</tr>
<tr>
<td>Versuch ‘attempt’</td>
<td>7</td>
<td>0.571</td>
<td>1.000</td>
<td>0.727</td>
</tr>
<tr>
<td>Ansicht ‘view’</td>
<td>42</td>
<td>0.471</td>
<td>0.889</td>
<td>0.615</td>
</tr>
<tr>
<td>Forderung ‘demand’</td>
<td>26</td>
<td>0.450</td>
<td>0.900</td>
<td>0.600</td>
</tr>
<tr>
<td>dies ‘this’</td>
<td>59</td>
<td>0.420</td>
<td>0.913</td>
<td>0.575</td>
</tr>
<tr>
<td>das ‘that’</td>
<td>191</td>
<td>0.430</td>
<td>0.829</td>
<td>0.566</td>
</tr>
<tr>
<td>Meinung ‘opinion’</td>
<td>30</td>
<td>0.364</td>
<td>1.000</td>
<td>0.533</td>
</tr>
<tr>
<td>Argument ‘argument’</td>
<td>5</td>
<td>0.333</td>
<td>1.000</td>
<td>0.500</td>
</tr>
</tbody>
</table>

**Table 5**: Classification performance of Naive Bayes classifier for particular anaphors
(5) **Eines** ist auch klar, und **dazu** stehen wir auch: *Kontrolle ist gut, Vertrauen in Kontrolle ist besser*. Aber **das** wird der Vorschlag dieser vier Staaten bei den Bürgerinnen und Bürgern in dieser Form nicht erreichen.

‘One thing is clear and we stand behind this too: Control is good, trust in control is better. But the proposal of these four states in this form won’t achieve this.’
Wir, die Vertreter der Bürger, der Menschen Europas, erreichen innerhalb der Institutionen etwas, und wir debattieren darüber kurz vor Mitternacht und werden ermahnt, wir sollen uns kurz fassen. Das ist ein Wiederholungsfall!

‘Within the institutions, we – the representatives of the citizens, of the people of Europe – are actually achieving something, and, as we debate it shortly before midnight, we are enjoined to keep things brief. This is not the first time this has happened.’
Get the source of this theme and the demo presentation from

github.com/matze/mtheme

The theme *itself* is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.