Anaphora resolution of the third-person pronoun in texts from narrow subject domains with grammatical errors and mistypings

Daniel Skatov, Sergey Liverko
Dictum Ltd., Nizhny Novgorod

Dialogue-21, Bekasovo, May 2011

Anaphora resolution

Of 3rd-person pronouns in opinions (about mobile phones) written in a free style, with grammatical errors, mistypings:

<LG KS660>
bought for bussines, very useful **it(he)** supports two SIM-cards. Nice and big displey, no dead zones on **it(him)**
</LG KS660>

Known methods (for «correct» well-formed texts, dossiers, …) aren't suitable.

General considerations

Antecedent A — a word which corresponds to pronoun P:

- A is in 1–2 «sentences» left from P,
- A и P — concorded by number and gender.
- A — closest to the left for P? Not always.

Bussines (mistypes are fixed)?
Phone (A is missed explicitly)?

<LG KS660>
bought for bussines, very useful **it(he)** supports two SIM-cards. Nice and big displey, no dead zones on **it(him)**
</LG KS660>

Display? Business? Mobile phone?

Additional effects

- **Chain**. 3 won't find 0 — it is more distant than in 2 «sentences». So we need to: 1→0, 2→1,3→2.

  <nokia 6730>
  Keyboard$_0$ is awesome! Sometimes **[it(she)]$_1$** seems unusual after samsung. **[its(her)]$_2$** keys are likely to be small. But it's not bothersome in respect to **[its(her)]$_3$** advantages)
  </nokia 6730>

- **«Implicit» antecedent**: a subject that is being discussed («mobile phone» = *), or it's manufacturer (#), and a resolving pronoun is abscent on the left side/just everywhere in text.

  <nokia n97>
  Bought it 19th of July, the day **[it(he)]{*}** was supposed to appear in shops!
  </nokia n97>

Task specificity

- We don’t mean only text singularities.
- The result of processing → decisions to be made by human, and no handheld postprocessing of the result.
- So there must be an output with high «reliability». We can even sacrifice the volume of our material (recall) to achieve high precision. **Less is more.** But not too less ☺
- Manually dropped out pronouns. How? →

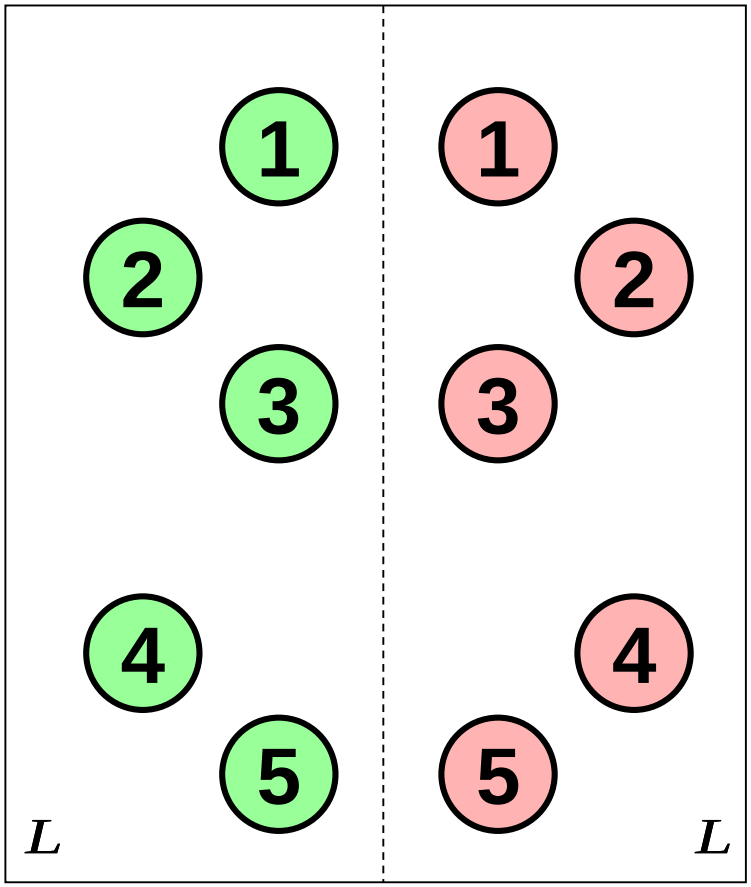
Approach to resolution

- For each pronoun
- Collect hypothesis' and rank them
- so that the 1st one is frequently correct:

  bought for bussines, very useful **[it] {* = 0.652166, business = 0.2371, NULL = 0.168611}** supports two SIM cards. Nice and big displey, no dead zones on **[it]{display = 0.466248, * = 0.284525, NULL = 0.0777368, business = 0.0101848}**

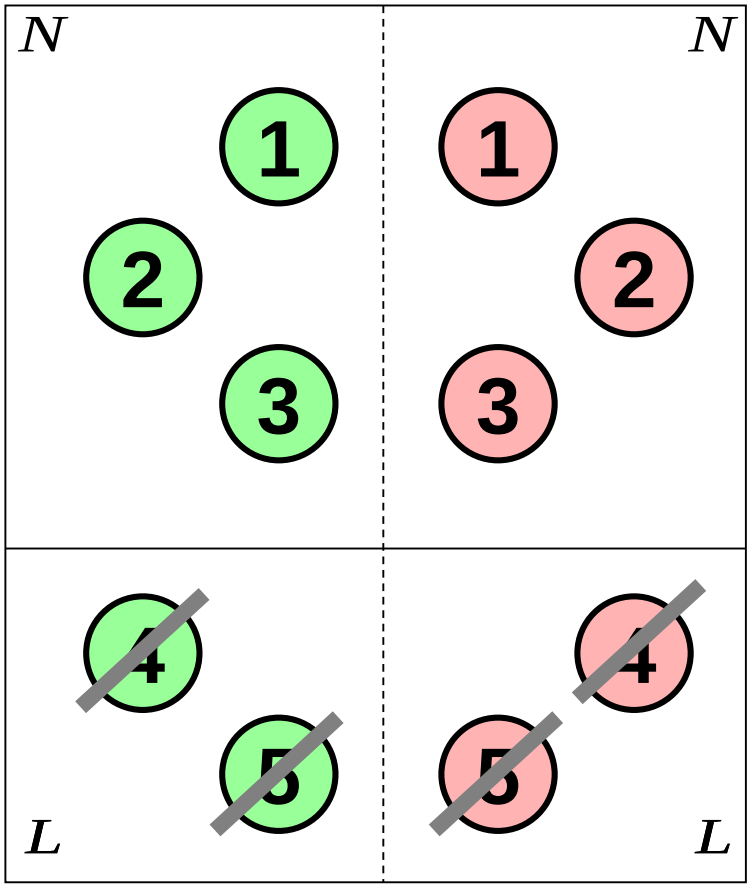
- Choose the 1st one as correct resolver.

Approach to evaluation



Two possible states for every pronoun: resolved correctly **(green)** or incorrectly **(red)**

Approach to evaluation

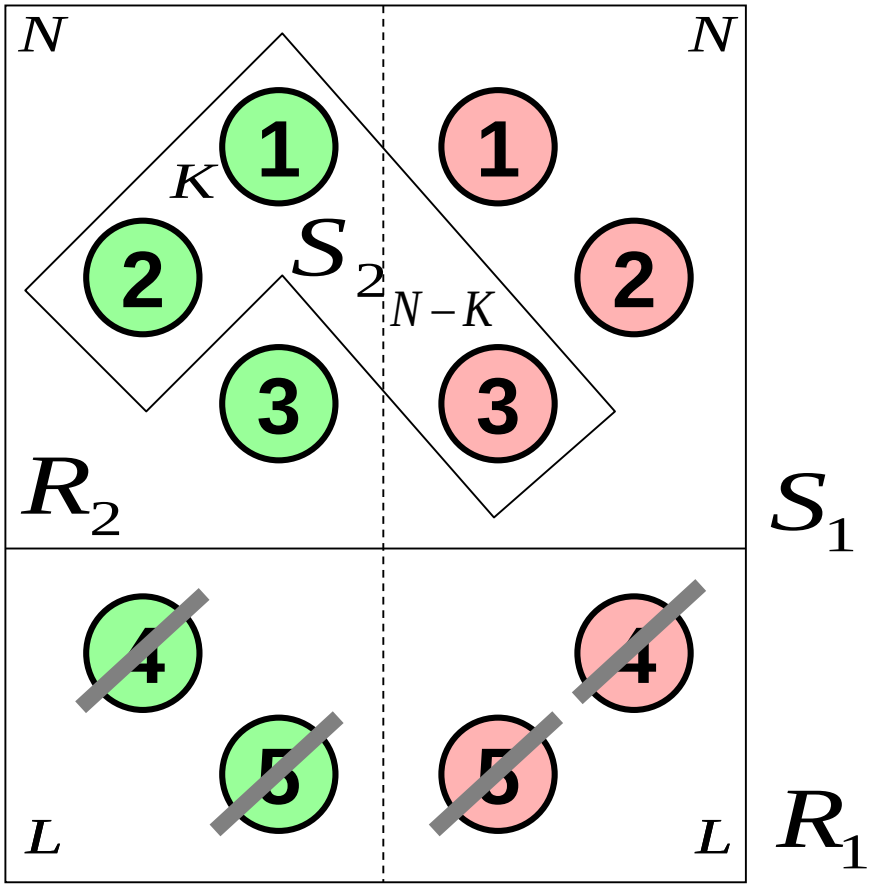


S_1

R_1

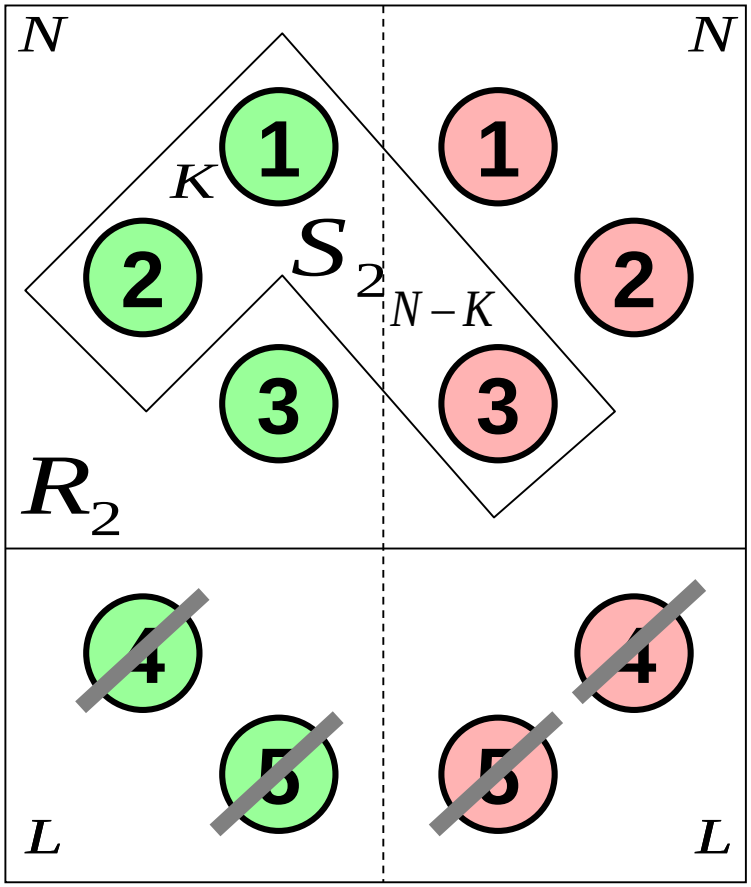
Two possible states for every pronoun: resolved correctly **(green)** or incorrectly **(red)**

Approach to evaluation



Two possible states for every pronoun: resolved correctly **(green)** or incorrectly **(red)**

Approach to evaluation



Two possible states for every pronoun: resolved correctly **(green)** or incorrectly **(red)**

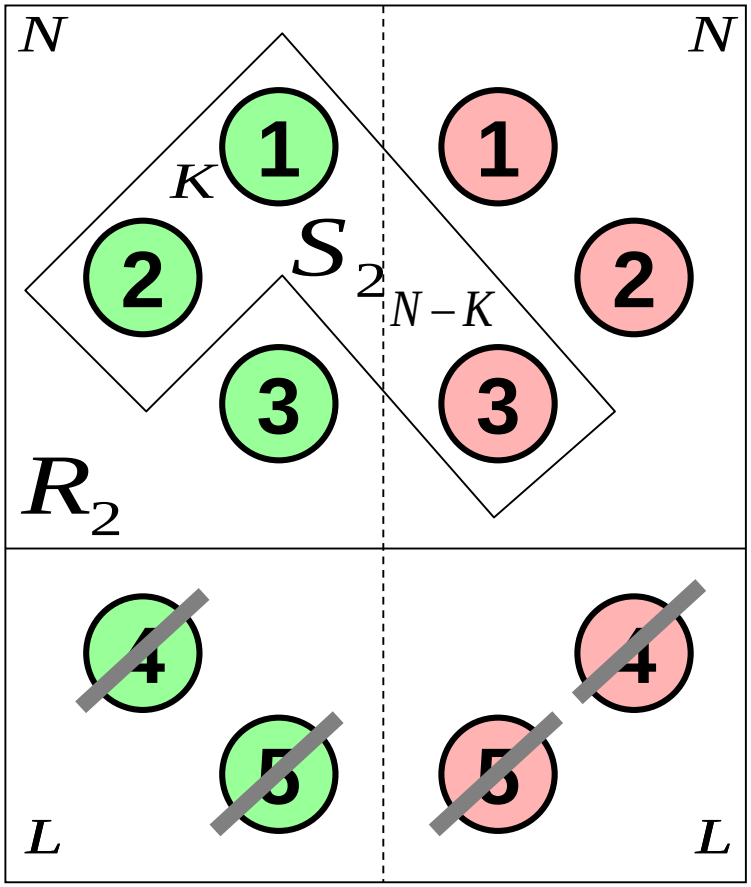
$$\text{Rec}_1 = \frac{|R_1 \cap S_1|}{|R_1|} = \frac{2N}{2L} = \frac{N}{L}$$

$$\text{Prec}_1 = \frac{|R_1 \cap S_1|}{|S_1|} = \frac{2N}{2N} = 1$$

$$\text{Rec}_2 = \frac{|R_2 \cap S_2|}{|R_2|} = \frac{K}{N}$$

$$\text{Prec}_2 = \frac{|R_2 \cap S_2|}{|S_2|} = \frac{K}{K+N-K} = \frac{K}{N}$$

Approach to evaluation



Two possible states for every pronoun: resolved correctly **(green)** or incorrectly **(red)**

$$\text{Rec}_1 = \frac{|R_1 \cap S_1|}{|R_1|} = \frac{2N}{2L} = \frac{N}{L}$$

$$\text{Prec}_1 = \frac{|R_1 \cap S_1|}{|S_1|} = \frac{2N}{2N} = 1$$

$$\text{Rec}_2 = \frac{|R_2 \cap S_2|}{|R_2|} = \frac{K}{N}$$

$$\text{Prec}_2 = \frac{|R_2 \cap S_2|}{|S_2|} = \frac{K}{K+N-K} = \frac{K}{N}$$

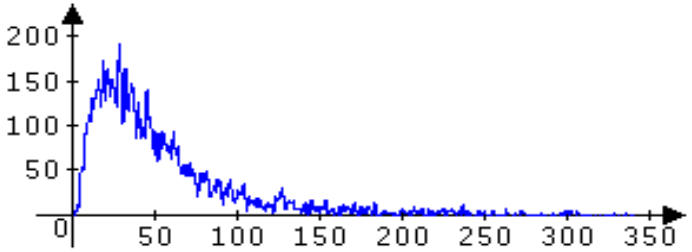
$$\longrightarrow \quad \boxed{(\text{Rec}, \text{Prec}) = \left(\frac{N}{L}, \frac{K}{N}\right)}$$

The corpora

- 3 Mb Win-1251 plain-text opinions, 3 equal parts = Neutral, Positive, Negative.
- Markup: only one resolver for each pronoun; is written near the pronoun in certain gender and number or implicit sign (***, #**).
- For explicit (which occur before the pronoun in text): a concrete location of the resolver in text (for further measurements).
- Special **NULL** for semantic ambiguity: *This mobile phone has a sensor screen. It's very inconvenient. (screen or phone?).* Hard to resolve even for human, so leave as is.

The corpora

- 8.3k opinions, 37k unique wordforms.
- Opinion length in words:
  15–35 are most frequent,
  54 in average, range 2–340.



- In sentences: 1–16 are most frequent, 4 in average, range 1–40.
- 6.2k 3rd-form personal pronoun in total, 4.5k male, 0.8 female, 0.7 plural.
- 50% of opinions contain at least one pronoun. 35% one, 10% two, 5% three and more. Not more than 9 pronouns in one opinion.

Feature space

- **IsVoc** — belonging of **A** to a domain dictionary;
- **Freq** — the number of mentionings of the given hypothesis **A** (in any form) to the left of **P**;
- **Dist** — the distance between **A** and **P** inside the text (measured in words);
- **IsVerb** — the presence of direct father in a form of verb in syntax tree for a «sentence» containing **A**;
- **NumNodes** — the number of nodes in a bush subordinate to **A**.

IsVoc is most difficult to prepare.

Lexicographical method

- Form vector of features for each hypothesis **`A`** for a given pronoun **`P`**;
- Sort those vectors lexicographically for **`P`**.
- Position of feature in vector = significance.
- *Evaluation:*

	With **`IsVoc`**	Without **`IsVoc`**
(Rec, Prec)	(93.7%, 51.9%)	(93.7%, 42.4%)

- Reason: the most significant feature kills others.

Move into vector space

- For every **P** with ***N*** hyphotesis' we have one **A** which «is antecedent» and ***N−1*** which are «is not antecedent». 2 classes for each **P**.
- Those are only for one **P**.
- What about ***M*** pronouns in one review? 2 classes for each **P**, ***2M*** classes for the entire opinion.
- Lets convolute: add a constant centroid to the end of the feature space vector, calculated over all the hypothesis' in the entire opinion. Then union «is antecedent» for all pronouns **P** in the opinion. The same for «is not antecedent». Now we have **2 classes for one opinion.**
- The trick gives a boost to quality performance.

Discriminant analysis (LD)

- While stepping away from heuristic-based lexicographical method, we tried to examine the nature of the feature space we built.
- The goal is to find out, are two clouds of data in vector space are separable somehow. Linear discriminant analysis builds a line that separates two clouds as best as possible. Then project two clouds on it — see how separable they are.
- Did it for classes «are antecedent» and inverse.
- Obtained promising results to go deeper.
- These results will also help us further.

Some interesting stats

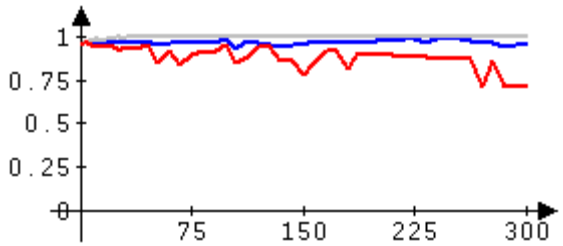
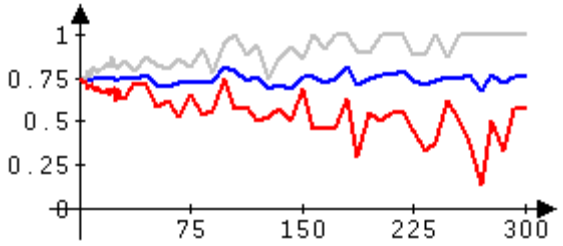
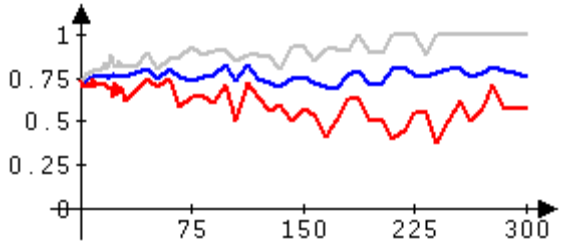
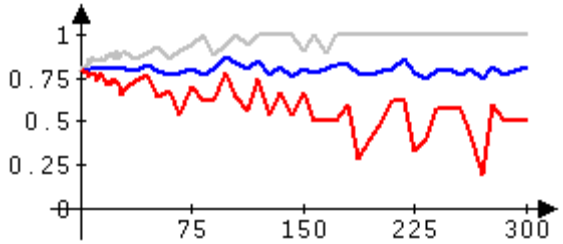
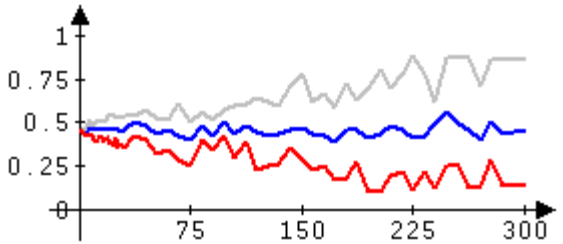
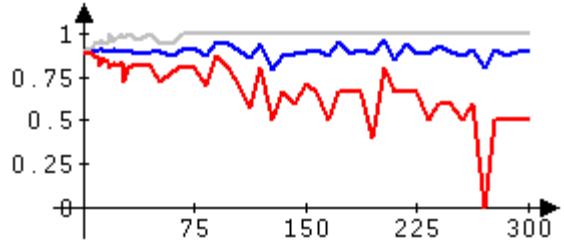
Feature	Coeff. in LD	Coeff in LD. for centroid
`IsVoc`	*9.3*	*– 1.1*
`Freq`	*– 21.5*	*– 1.6*
`Dist`	*– 10.6*	*0.1*
`HasVerb`	*– 7*	*35.8*
`NumNodes`	*– 0.5*	*18.9*

SVM-based method

- Use C-SVM with polynomial kernel.
- Combination with LD to filter pronouns.
- q-fold validation for 4k opinions, q=1..300.

Method	Recall	Precision
A	97.3%	74.2%
B	75.4%	80.7%
C	45.6%	90.3%

SVM-based method

	Recall	Precision
A		
B		
C		

Where in this world are typos?

- For each noun, all grammar values of every spelling correction variant are united into one list of grammar values.
- Those grammar values are taken into account in syntax parser. This parser can help to choose a proper grammar value for words with typos.
- Gr.V. lists sometimes large. Additional filtering is reasonable. *Пора (it's time, but not pore), мум (here, but not mulberry tree), уж (really — grass snake), потом (later — sweat), рада (happy — Rada, unit), том (that — tome), ком (who — lump)* etc.

Thanks for your attention!

Questions?