University of Zurich[UZH]

Institute of Computational Linguistics

Non-negative Matrix Factorisation-based Verb Semantics for 3rd Person Pronoun Resolution

Don Tuggener, Manfred Klenner
tuggener@cl.uzh.ch

20. September 2012

Outline

- Distributional Verb Semantics
- I.e. Selectional Preferences of Verbs
- Application: 3rd Person Pronoun Resolution (German)

Example

Judging from the Americana in Haruki Murakami's "A Wild Sheep Chase" Kodansha, 320 pages, $ 18.95 , baby boomers on both sides of the Pacific have a lot in common. Although in Japan, the novel's texture is almost entirely Western, especially American. Characters drink Salty Dogs, whistle "Johnny B. Goode" and watch Bugs Bunny reruns.
They read Mickey Spillane and talk about Groucho and Harpo.

Example

Judging from the Americana in Haruki Murakami's "A Wild Sheep Chase" Kodansha, 320 pages, $ 18.95 , baby boomers on both sides of the Pacific have a lot in common. Although in Japan, the novel's texture is almost entirely Western, especially American. Characters drink Salty Dogs, whistle "Johnny B. Goode" and watch Bugs Bunny reruns.
They read Mickey Spillane and talk about Groucho and Harpo.
→ *Morphologically compatible, possible antecedents (Number, plural NPs)*
→ *Which one is it?*

Example

Judging from the Americana in Haruki Murakami's "A Wild Sheep Chase" Kodansha, 320 pages, $ 18.95 , baby boomers on both sides of the Pacific have a lot in common. Although in Japan, the novel's texture is almost entirely Western, especially American. Characters drink Salty Dogs, whistle "Johnny B. Goode" and watch Bugs Bunny reruns.
They read Mickey Spillane and talk about Groucho and Harpo.
→ *Morphologically compatible, possible antecedents (Number, plural NPs)*
→ *Which one is it? How do we know?*

Example

Judging from the Americana in Haruki Murakami's "A Wild Sheep Chase" Kodansha, 320 pages, $ 18.95 , baby boomers on both sides of the Pacific have a lot in common. Although in Japan, the novel's texture is almost entirely Western, especially American. Characters drink Salty Dogs, whistle "Johnny B. Goode" and watch Bugs Bunny reruns.
They read Mickey Spillane and talk about Groucho and Harpo.
→ *Morphologically compatible, possible antecedents (Number, plural NPs)*
→ *Which one is it? How do we know?*
→ *They must be something that can read and talk*

Motivation

- Morphosyntactic agreement/features: Often many compatible antecedent candidates (see real world example above: 4 Sentences, 1 pronoun, 8 possible referents!)
- Rules or ML classifiers for antecedent selection (Grammatical functions, distance etc., e.g.: choose nearest subject)

Our approach

- Make use of discourse features, i.e. verbs and their selectional *preferences*
- $\{Banker, Poet\} \rightarrow$ "He loves flowers".
- Both $Banker$ and $Poet$ are morphologically compatibel and "valid"
- How likely is it that someone like a $Banker|Poet$ does something like $loving$ and that the object of this is something like a $flower$?
- $p(banker_{subj}|loving, flower_{obj})$ vs $p(poet_{subj}|loving, flower_{obj})$
- Higher p yields antecedent of He

Method

- Distributional model of verb selectional preferences
- Co-occurrence matrices $subj \times verb$ and $subj \times obj$
- E.g. rows consist of nouns, columns are verbs

Verbs / *Nouns*	love	sell	...
poet	11	4	...
banker	3	17	...
...	...	...	...

Method II

Nouns \ *Verbs*	love	sell	...
poet	11	4	...
banker	3	17	...
...	...	...	...

- Sparsisty a big problem, cannot expect to find all possible filler objects for the verbs even in huge corpora
- Especially true for German: Noun compounding (Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz; law of delegating beef labeling supervision duties)
- Need (semantic) smoothing, i.e. reduction of the number of zero-value cells / unseen but possible combinations
- (Non-negative) Matrix Factorisation

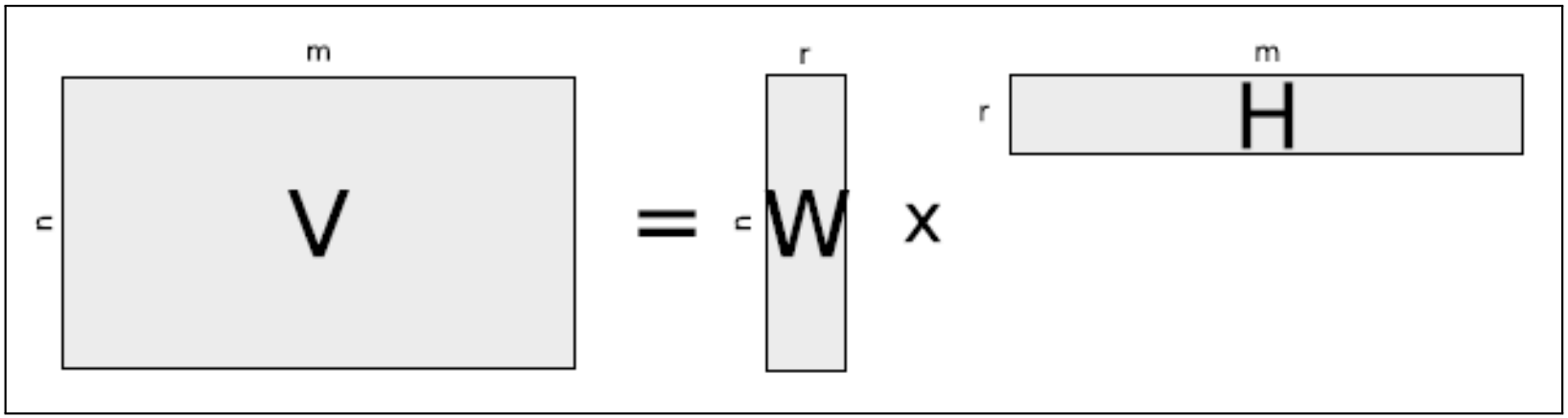
Non-Negative Matrix Factorisation

- Factorisation technique similar to Latent Semantic Analysis (LSA or probabilistic LSA) and k-means clustering
- Decompose original matrix into two smaller matrices with r dimensions

$$V_{n\times m} \approx W_{n\times r} H_{r\times m}$$

- Approximates V by minimising an error/objective function (Kullback-Leibler Divergence suited for language data)
- Dimension reduction hypothesized to yield latent dimensions/classes
- E.g. *like* and *love* are similar as they choose similar subjects/objects, they form/belong to a latent class
- Only non-negative values, enables probabilistic interpretation

Non-Negative Matrix Factorisation II



Van de Cruys (2010)

- W and H are initialized randomly
- Update rules for W and H, alternatively applied during iteration (i.e. approximation), until max iterations or threshold in the change of the objective function

Non-Negative Matrix Factorisation III

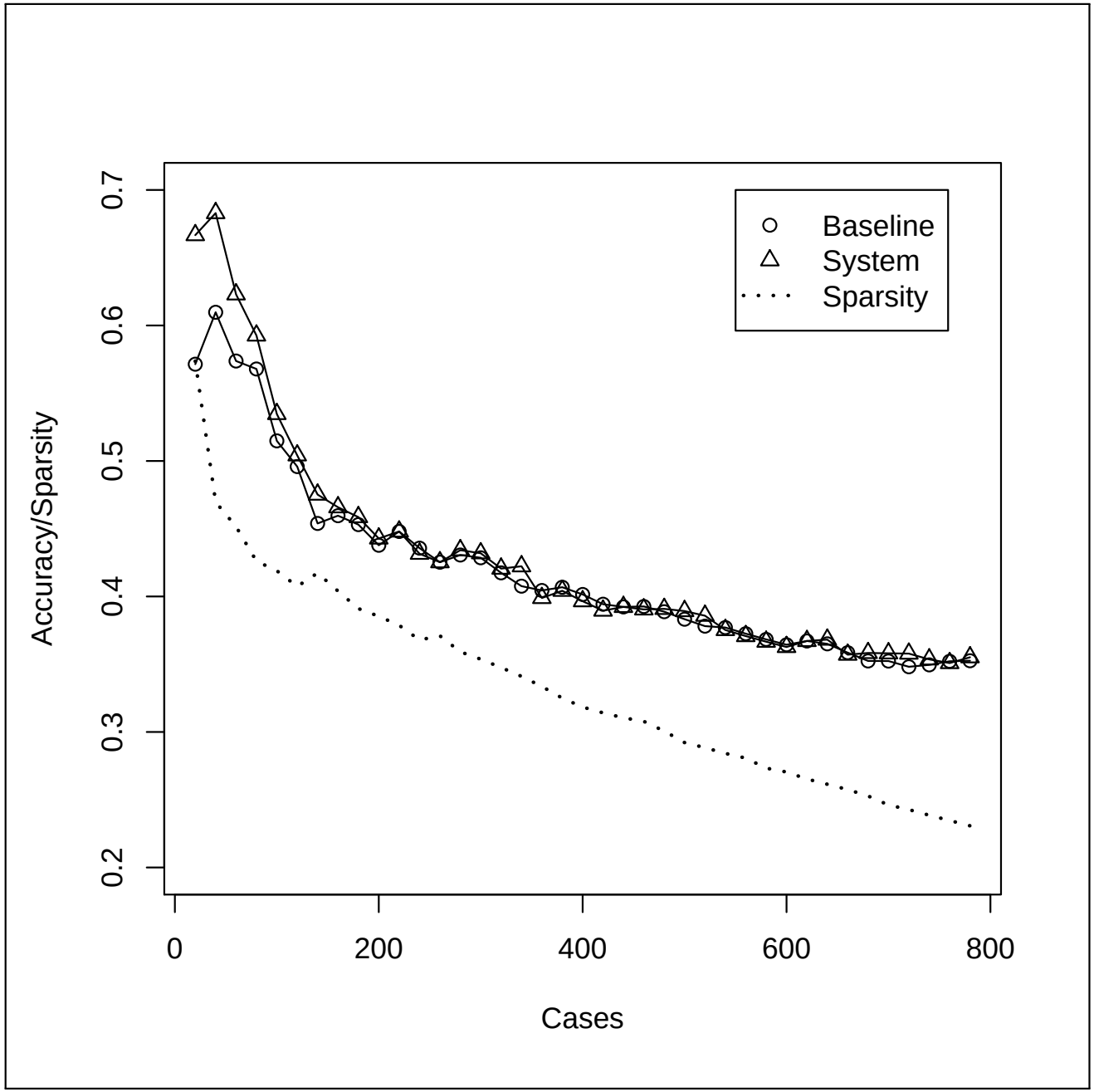
- V_{approx} generated by matrix multiplication of $W_{n \times r}$ and $H_{r \times m}$
- $V_{approx} \neq V$, because of dimension reduction, number of iteration
- Number of iterations and dimension and latent dimensions ($W_{n \times r}$ and $H_{r \times m}$) not that important (to us)
- We are interested in the smoothing effect, i.e. the reduction of zero value cells, in V_{approx} compared to V
- V_{approx} is based on latent classes $\rightarrow$ kind of semantic smoothing, adapting the bare corpus-based frequencies to their hidden inter-dependencies

Data

- TübaD/Z German Treebank, coreference annotation
- Cases of pronouns governed by transitive verbs (which occur at least 10 times)
- Coreference system extracted antecedent candidates
- 109 verbs in 780 test cases, 3666 antecedent candidates → Basis for matrix construction
- Dewac as corpus for instantiation of V
- Dependency parsing, NE classification
- c.a. 5 Mio subj-verb-obj tripels extracted
- Two matrices
  - subj-verb → $p(subj|verb)$
  - subj-obj → $p(subj|obj)$

Evaluation

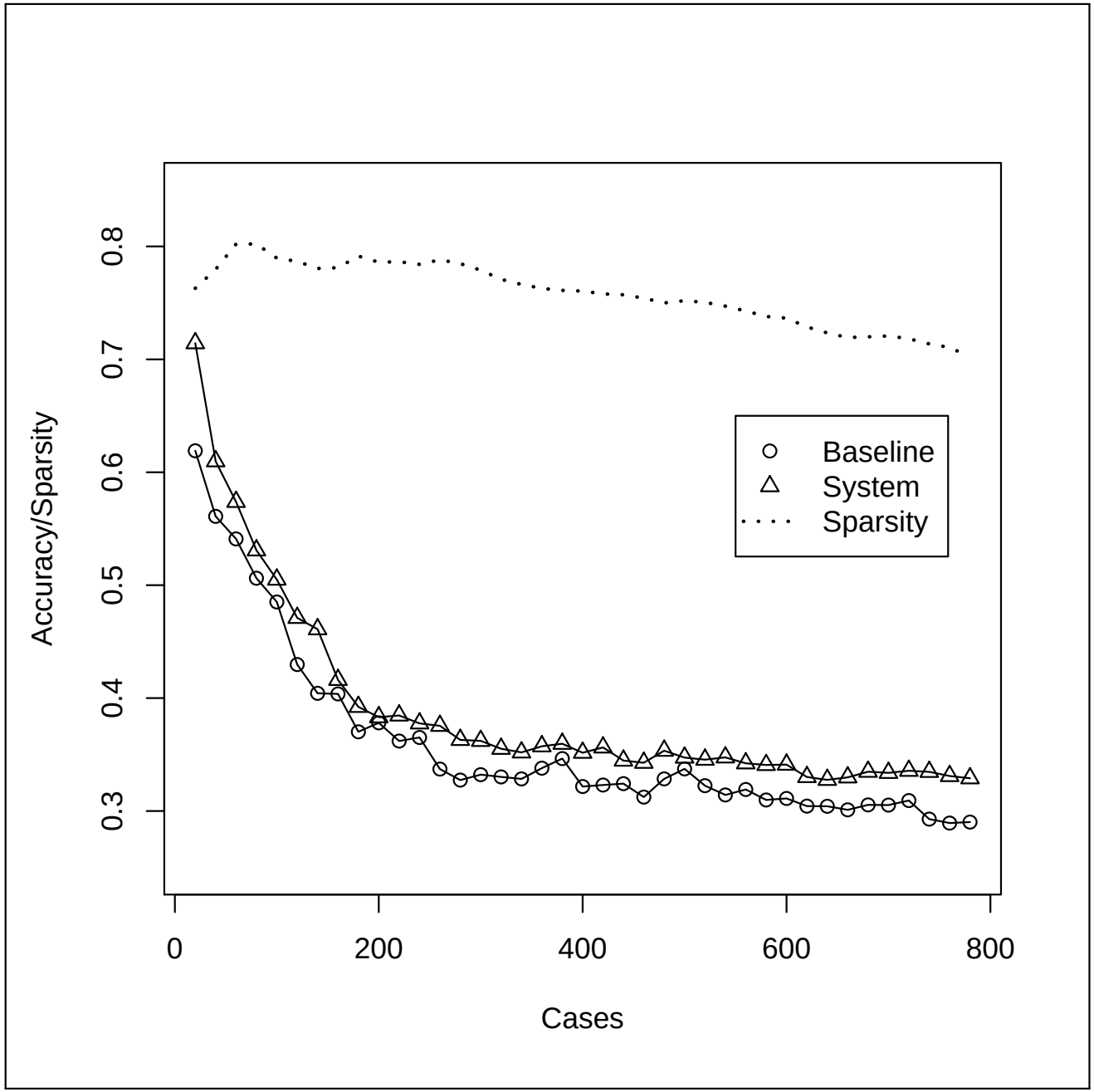
- Compare antecedent selection based on p from original V and smoothed V_{approx}
- Test data selection: Frequency based selection. Test cases (verb and antecedent candidates) with highest frequency for each noun in Dewac. → should be nouns for which V is not too sparse
- Evaluation: Start with "good" cases (highly frequent nouns in Dewac), add cases with less frequent nouns incrementally
- Accuracy: $\frac{Correctly\ resolved\ Pronouns}{All\ Pronouns}$
- Matrix sparsity: Percentage of non-zero value cells



Smoothing the Subject-Verb Matrix

Smoothing the Subject-Verb Matrix

- Original and smoothed matrices both perform better than chance, which is about 20% given 4.7 antecedent candidates per pronoun/verb
- The 20 “best” cases (with highly frequent antecedent candidates in the Dewac) start with almost 70% Accuracy (that’s pretty good!)
- Accuracy drops to 30% given all test cases (that’s pretty bad!)
- Performance difference between V an V_{approx} very marginal after 100 best cases!
- That’s also when matrix sprasity of V drops below 50%!



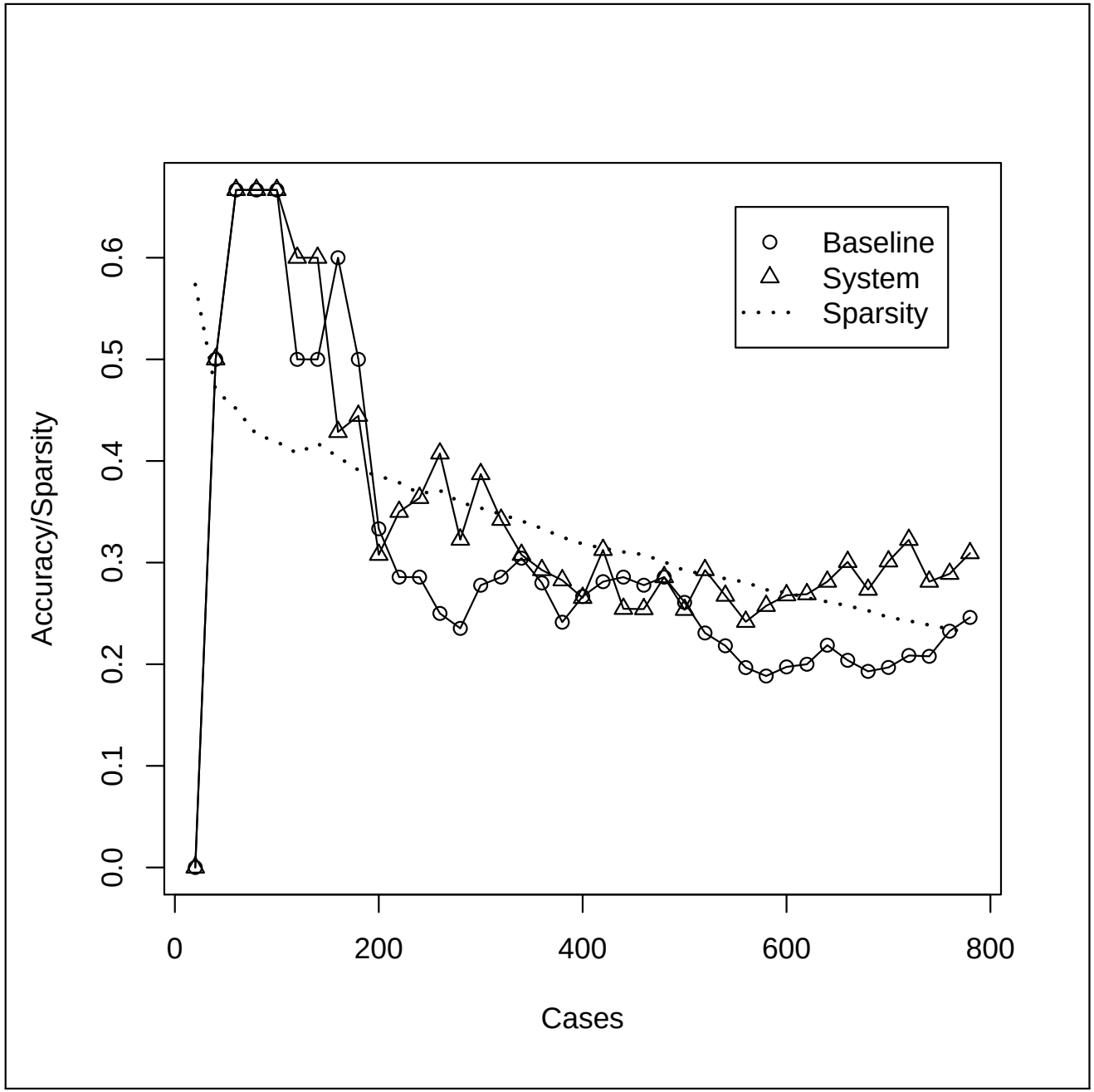
Smoothing the Subject-Obj Matrix

Smoothing the Subject-Obj Matrix

- Similar findings to subj-verb setting
- Accuracy values generally lower
- Matrix sparsity not that problematic, although counts less frequent with increasing test cases, impairing the model
- Combining the models $p(subj|verb) * p(subj|obj)$ didn't lead to improvements!

Smoothing special cases

- Special cases where the antecedent candidate with highest p in V_{approx} has zero-value in V
- How often is this augmentation of the zero-value candidate correct? I.e. yields the true antecedent?
- Accuracy only of these cases



Smoothing special cases

Smoothing special cases

- For the rather non-sparse cases the NMF approach is not that helpful, accuracy $V_{approx} = V$
- I.e. almost no sparse (true) antecedent candidates in the "best" 100 cases
- When sparsity increases, the accuracy difference between the V and V_{approx} increases and becomes evident and significant
- NMF actually helps here

Conclusion & Outlook

- Both methods perform significantly better than chance, it is reasonable to do it!
- If the co-occurrency matrix V contains less than ca. 50% zero-value cells, the approach works, i.e. factorisation-based V_{approx} yields better results
  - Decomposition of compound nouns (German!); (Rindfleischetikettierungsüberwachungsaufgabenübertragungsge → Gesetz (Law))
    Class abstraction (sculptor → artist; level of abstraction? VerbNet abstractions applicable to German?)
- Use this in a real Coreference/Pronoun resolution system
  - no point in performing pronoun resolution using only this model
  - derive features/hard contraints for antecedent filtering in a real system

University of Zurich[UZH]

Institute of Computational Linguistics

Thank you! :)