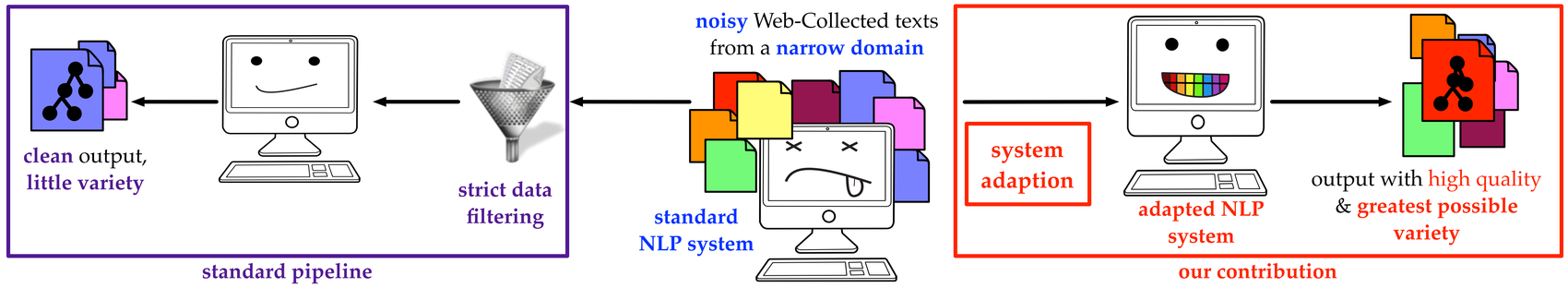
Robust Processing of Noisy Web-collected Data

Jelke Bloem*, Michaela Regneri**, Stefan Thater**
*University of Groningen, **Saarland University
j.bloem.3@student.rug.nl, regneri@coli.uni-saarland.de, stth@coli.uni-saarland.de

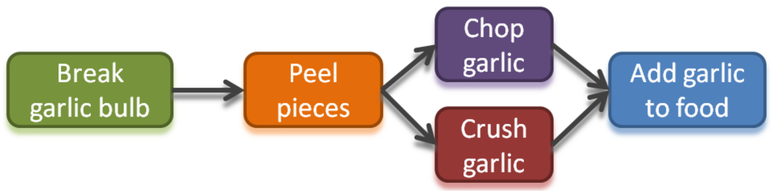


Introduction

Crowdsourced linguistic data is often **noisy** in some way. Instead of throwing out unsuitable data, we show that **adaptation of NLP tools** to the data allows automatic processing with good accuracy for a narrow but unknown domain, **keeping more information**. This poster demonstrates the concept for **spell-checking** and **pronoun resolution** tools.

Event Sequence Descriptions

Our data is a web-collected data set of **descriptions of kitchen tasks**, gathered for script mining[2]. A script describes such a scenario:



Example of one web-collected kitchen task description:

1. first strip of the papery skin of the bulb
2. ease out as many intact cloves as possible
3. chop them finely if you want a stronger taste
4. chope them coarsely if you want a weaker taste
5. crushed garlic is the strongest taste

Script Mining Task

- Many web-collected descriptions are used to create a single model script
- Different descriptions should be matched:
  chop them = **chip the garlic up**.
- This requires preprocessing:
  - Spelling correction: Makes for better input for standard applications or dictionaries.
  - Pronoun resolution: To know what the object in '*chop them*' is.

Spelling Correction

To adapt the spelling correction tool to our data, we use an unmodified general tool but add domain-sensitive heuristics to select detections and corrections.

General use: GNU Aspell

- Uses a general dictionary (which may lack kitchen-specific terms or names)
- This leads to false corrections of correct words such as:

```
deglaze -> deg laze
microplane grater -> micro plane grater
ziploc -> zoology
```

Modified Aspell

We also consider the domain context (**slice bread**):

- If a word occurs in at least 3 other descriptions, the system accepts it
- Only accept corrections that occur in at least 1 other description, preventing:
  `bord -> bird (board), loaft -> loft (loaf)`
- Split only if the resulting words occur in another description, preventing:
  `sandwhich -> sand which (sandwich)`

The intended word is in parentheses. These off-topic corrections would be plausible in different domains than '**slice bread**'.

Evaluation

Method	Precision	False Positives	True Precision	Sem. Precision	Corrections
Aspell	0.43	0.28	0.57	0.58	162
Enhanced Aspell	**0.58**	0.29	**0.79**	**0.76**	150

Evaluation based on manual judgement of the corrections made by the spell-checkers. Bolded results indicate an improvement over the baseline.

References

[1] E. Charniak and M. Elsner. EM works for pronoun anaphora resolution. In *Proceedings of EACL*, pages 148–156. Association for Computational Linguistics, 2009.

[2] Marcus Rohrbach, Michaela Regneri, Micha Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision - ECCV 2012 : 12th European Conference on Computer Vision*, volume 2012 of *Lecture Notes in Computer Science*, Firenze, Italy, October 2012. Springer, Springer.

[3] Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. Word meaning in context: A simple and effective vector model. In *Proc. of IJCNLP 2011*, 2011.

Pronoun Resolution

Slice bread

1. put the bread on the cutting board
2. hold onto the middle and cut the heel off
3. keep slicing at whatever thickness you want
4. keep moving your holding point so you can slice it until the end

Which candidate referent does the pronoun 'it' refer to?

General use: EM-based [1]

- General, openly available pronoun resolver
- Relies on **grammatical** features

General use: Vector space model [3]

- Model for selectional preference of verbs, trained on a large corpus
- Compares meaning vectors for different candidate antecedents

Context association: Odds ratio

- Selectional preference model, computed only on our data set
- Compares probability of each candidate occurring with the main verb
  `slice cutting board`, `slice knife`, `slice bread`
  Which is more likely in this domain?

Evaluation

Model	Correct
Vector space model	0.544
EM	0.175
Odds ratio	**0.631**

Our *odds ratio* model compared to two baselines.

- We cannot count on grammar features due to unusual writing style and noise
- Our simpler but domain-specific method method outperforms a complex general model

Conclusions

- Adapt tools to properties of your data set (i.e. parallelism) for better preprocessing
- State-of-the art performance can be reached with simple methods and heuristics
- This approach preserves more crowdsourced data than traditional filtering
- The specific methods we used could be refined further