A generalized method for iterative error mining in parsing results

Daniel de Kok, Jianqiang Ma, Gertjan van Noord

GEAF workshop 2009 - August 6, 2009

What is error mining?

Two common types of parsing problems:

- Incorrect parse, e.g. by incorrect disambiguation
- Incomplete parse: no analysis spanning the full sentence could be found, usually due to missing lexicon items or an incomplete grammar

What is error mining? (2)

The basic idea:

- Parse a (unannotated) corpus
- Extract parsable and unparsable sentences
- Find n-grams that occur in unparsable sentences, but not in parsable sentences
- Assign some score to n-grams

Overview

- Previous work
- N-gram expansion
- Sparseness correction
- Finding patterns

Van Noord (2004)

- Ratio-based mining, the suspicion of an n-gram is defined to be

$$S(w_i..w_j) = \frac{C(w_i...w_j|error)}{C(w_i...w_j)} \tag{1}$$

- Suspicion by accident: one or just a few forms are responsible for most parsing failures, but all forms occurring in an unparsable sentence take blame

Sagot and de la Clergerie (2006)

Iterative error mining method wherein:

- If a form occurs in a parsable sentence, it becomes less likely that it is to blame.
- The suspicion of a forms should depend on the company it keeps
- A form observed in a short sentence is initially more suspicious than a form observed in a longer sentence.

Usefulness of n-grams

- Problem of the miner described by Sagot and de la Clergerie: only mines unigrams and bigrams
- Prior experience with the Van Noord (2006) miner show n-grams are very useful:
- *de* (*the*), *eerste* (*first*), *beste* (*best*) had low suspicions
- *eerste de beste* (as occuring in *de eerste de beste*) had a very high suspicion

Mining of n-grams

- Blindly adding n-grams as forms distorts mining, consider the sequence *A B C* where *B* only occurs in unparsable sentences
- Adding all n-grams for a larger *n* is expensive

Preprocessor

- Iterate through a sentence by unigram
- Try to extend each unigram stepwise, where an extension is allowed if the (ratio-based) suspicion of an n+1-gram is higher than both of its n-grams:

$$S(i..j) > S(i..j-1) \tag{2}$$

$$S(i..j) > S(i+1..j) \tag{3}$$

- The sentence is represented by an n-gram for every sentence position, potentially extending to the end of the sentence.

Data sparseness

- To handle data sparseness, we added a factor that is dependent on the form frequency, expansion only happens when

$$S(i..j) > S(i..j-1) \cdot extFactor \tag{4}$$

$$S(i..j) > S(i+1..j) \cdot extFactor \tag{5}$$

- This factor requires that the extended n-gram is frequent or much more suspicious

Evaluation methodology

- We want to improve the coverage of a grammar: we are interested in seeing forms with many unparsable sentences first (recall)
- We are interested in forms that primarily occur in unparsable sentences (precision)
- Combined: f-score, or in our case f0.5-score, placing more emphasis on precision

Quantitative evaluation material

Quantitative testing was performed on the Dutch Wikipedia corpus, that was parsed with the wide-coverage Alpino parser.

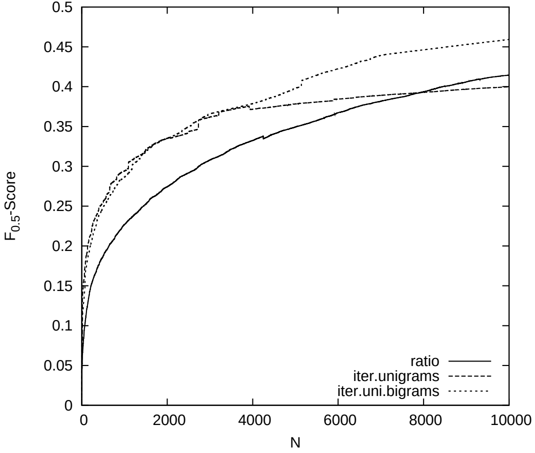
- 109 million words
- 7 million sentences
- 8.4% of the sentences were unparsable

Qualitative evaluation material

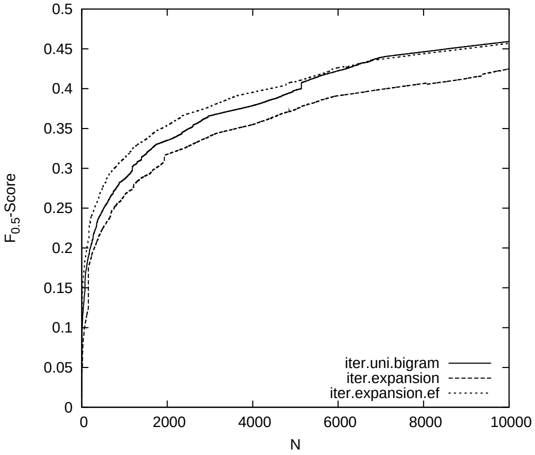
Qualitative evaluation was performed on the Flemish Mediargus corpus, that was also parsed with the Dutch Alpino parser:

- 1.1 billion words
- 67 million sentences
- 9.2% unparsable.

Results (iterative mining)



Results (expansion)



Flemish expressions

- Telkens hij [Everytime he]
- (had er AMOUNT) voor veil [(had AMOUNT) for sale]
- (om de muren) van op te lopen [to get terribly annoyed by]
- Ik durf zeggen dat [I dare to say that]
- op punt stellen [to fix/correct something]
- de daver (op het lijf) [shocked]
- (op) de tippen (van zijn tenen) [being very careful]
- ben fier dat [am proud of]
- Nog voor halfweg [still before halfway]
- (om duimen en vingers) van af te likken [delicious]

Long n-grams

- Het stond in de sterren geschreven dat NAME
- zowat de helft van de [...]
- er zo goed als zeker van dat
- laat ons hopen dat het/dit lukt

Pattern expansion

- Expand the notion of forms to mixed patterns, consisting of e.g. words, part of speech tags or lemmas
- Same procedure for expansion, but with additional considerations. For instance

$$S(w1, w2, t3) > S(w1, w2) \cdot extFactor \tag{6}$$

$$S(w1, w2, t3) > S(w2, t3) \cdot extFactor \tag{7}$$

- Prefer more abstract elements first

Implementation

- Suffix arrays inadequate to calculate pattern frequencies
- Hash table for each type of information, containing the set of corpus indices as values.
- We can now calculate the frequency of the pattern *i..j*:

$$I_{i..j} = (I_{i..j-1} + 1) \cap I_j \tag{8}$$

$$f_{i..j} = |I_{i..j}| \tag{9}$$

Evaluation material

- Unparsable and parsable sentences, randomly selected from the Mediargus corpus
- POS tagging was performed with the Citar HMM tagger, trained on the Dutch Eindhoven corpus
- Pattern expansion on words and POS tags

Pattern examples

- *doorheen N*: We reden met de auto doorheen Frankrijk (*We drove by car through France*)
- *wegens Prep Adj*: *Dat idee werd snel opgeborgen wegens te duur* (*That idea became soon archived because of too expensive*)

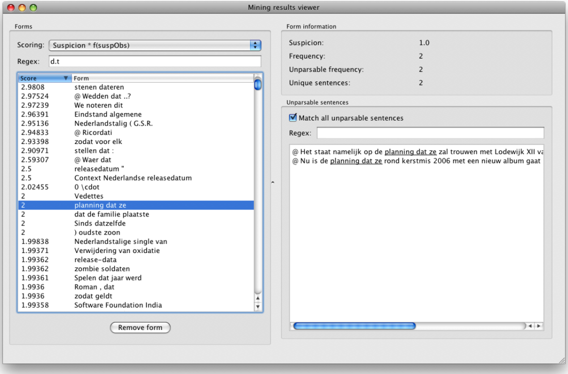
Improvement through pattern expansion

- More adequate/abstract descriptions of errors
- Consolidation, for instance 120 different problematic n-grams starting with *wegens* could be represented by the single pattern *wegens Prep Adj*

Conclusions

- Expanding to n-grams can give useful patterns
- Correction for sparseness is required to avoid making patterns too specific
- We provide method for quantitative evaluation of error mining
- Allowing for other information, such as POS tags gives rise to more general patterns

Software



Scoring: Suspicion * f(suspObs)

Regex: d.t

Score	Form
2.9808	stenen dateren
2.97524	@ Wedden dat ..?
2.97239	We noteren dit
2.96391	Eindstand algemene
2.95136	Nederlandstalig (G.S.R.
2.94833	@ Ricordati
2.93398	zodat voor elk
2.90971	stellen dat :
2.59307	@ Waer dat
2.5	releasedatum "
2.5	Context Nederlandse releasedatum
2.02455	0 \cdot
2	Vedettes
2	planning dat ze
2	dat de familie plaatste
2	Sinds datzelfde
2	) oudste zoon
1.99838	Nederlandstalige single van
1.99371	Verwijdering van oxidatie
1.99362	release-data
1.99362	zombie soldaten
1.99361	Spelen dat jaar werd
1.9936	Roman , dat
1.9936	zodat geldt
1.99358	Software Foundation India

Suspicion:	1.0
Frequency:	2
Unparsable frequency:	2
Unique sentences:	2

☑ Match all unparsable sentences

Regex:

@ Het staat namelijk op de planning dat ze zal trouwen met Lodewijk XII va

@ Nu is de planning dat ze rond kerstmis 2006 met een nieuw album gaat

Fast iterative miner, extensions, mining viewer: http://www.let.rug.nl/dekok/errormining/