Machine Learning for Rhetorical Figure Detection: More Chiasmus with Less Annotation

Marie Dubremetz & Joakim Nivre

UPPSALA
UNIVERSITY
SWEDEN

Nodalida - May 2017

Introduction

Introduction 

Chiasmus/Antimetabole: Traditional Definition

A rhetorical figure in which two words are repeated in reverse order.

Introduction 

Chiasmus/Antimetabole: Traditional Definition

A rhetorical figure in which two words are repeated in reverse order.

Example

Twist **facts** to suit **theories**,

not **theories** to suit **facts**.

Why Finding Chiasmus? 

chiasmus in corpus.txt

Results:

1. **First** should be **last** and **last** should be **first**
2. **Begining** of the **end**, the **end** of the **begining**
3. **Failing** to **prepare** is **preparing** to **fail**

...

99,999. **I** like beer from **time** to **time** but **I** prefer wine

- Practical: Text mining of master pieces and literature
- Linguistic: Improve our general knowledge of the figure?
- Proof of concept: If we can make it for chiasmus, you can hope to make it for more devices.

Why Finding Chiasmus? 

chiasmus in corpus.txt

Results:

1. **First** should be **last** and **last** should be **first**
2. **Begining** of the **end**, the **end** of the **begining**
3. **Failing** to **prepare** is **preparing** to **fail**

...

99,999. **I** like beer from **time** to **time** but **I** prefer wine

- Practical: Text mining of master pieces and literature
- Linguistic: Improve our general knowledge of the figure?
- Proof of concept: If we can make it for chiasmus, you can hope to make it for more devices.

Why Finding Chiasmus? 

chiasmus in corpus.txt

Results:

1. **First** should be **last** and **last** should be **first**
2. **Begining** of the **end**, the **end** of the **begining**
3. **Failing** to **prepare** is **preparing** to **fail**

...

99,999. **I** like beer from **time** to **time** but **I** prefer wine

- Practical: Text mining of master pieces and literature
- Linguistic: Improve our general knowledge of the figure?
- Proof of concept: If we can make it for chiasmus, you can hope to make it for more devices.

Why Finding Chiasmus? 

chiasmus in corpus.txt

Results:

1. **First** should be **last** and **last** should be **first**
2. **Begining** of the **end**, the **end** of the **begining**
3. **Failing** to **prepare** is **preparing** to **fail**

...

99,999. **I** like beer from **time** to **time** but **I** prefer wine

- Practical: Text mining of master pieces and literature
- Linguistic: Improve our general knowledge of the figure?
- Proof of concept: If we can make it for chiasmus, you can hope to make it for more devices.

State of the Art

State of the Art

The research on chiasmus

- Gawryjolek [2009]: Extract every double pair of words with reverse order without exception
  Chuck **Norris** does not fear **death**, **death** fears Chuck **Norris**
  - 100% recall
  - Very low precision ($< 1\%$)
- Hromada [2011]: Identify not two but three pairs of reverted words
  Love makes time pass, **time makes love** pass.
  - Very high precision
  - But low recall

State of the Art 

Problem

There are criss-cross patterns that are not chiasmi such as:

'**I** like beer from **time** to **time** but **I** prefer wine'

They are frequent but chiasmi are rare.
Consequence: the annotation task was endless, there was no corpus available.

State of the Art 

Problem

There are criss-cross patterns that are not chiasmi such as:

'**I** like beer from **time** to **time** but **I** prefer wine'

They are frequent but chiasmi are rare.
Consequence: the annotation task was endless, there was no corpus available.

State of the Art 

Problem

There are criss-cross patterns that are not chiasmi such as:

'**I** like beer from **time** to **time** but **I** prefer wine'

They are frequent but chiasmi are rare.
Consequence: the annotation task was endless, there was no corpus available.

State of the Art 

Problem

There are criss-cross patterns that are not chiasmi such as:

'**I** like beer from **time** to **time** but **I** prefer wine'

They are frequent but chiasmi are rare.
Consequence: the annotation task was endless, there was no corpus available.

This is the problem of the needle in the haystack!

(Re-)Defining the Task

Hard constraints are not a solution

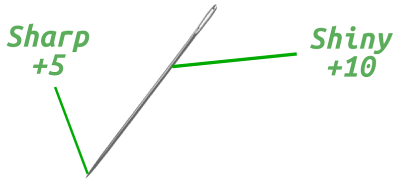
(Re-)Defining the Task 

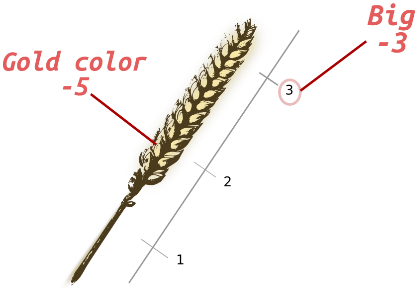
Hard constraints are not a solution

...Why not outputting chiasmi in a sorted manner?

Dubremetz & Nivre [2015]

Features





Our Model 

A standard linear model

So far 22 features have been successfully tested they encode: stopwords, lexical clues, ngram similarity, size, tag and parsing features

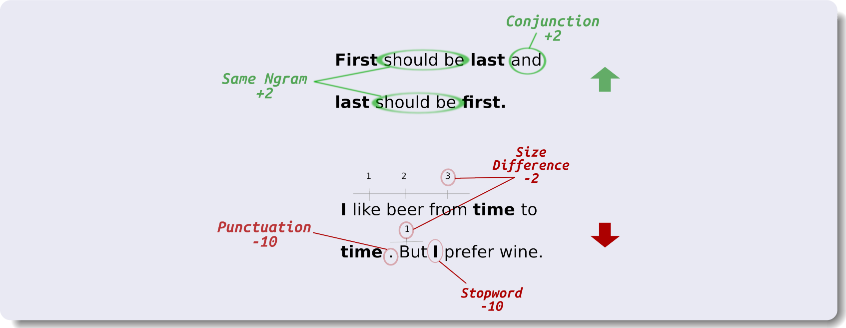
Our Model 

A standard linear model

So far 22 features have been successfully tested they encode:

- #punct
- #softPunct
- #centralPunct
- isInStopListA
- isInStopListB
- #mainRep
- #diffSize
- #toksInBC
- exactMatch
- #sameTok
- simScore
- #sameBigram
- #sameTrigram
- #sameCont
- hasConj
- hasNeg
- hasTo
- sameTag
- #sameDep W_a W'_b
- #sameDep W_b W'_a
- #sameDep W_a W'_a
- #sameDep W_b W'_b

An Example of Features



How our algorithm sorts criss-cross patterns:
5 representative examples of our 22 features

How Do We Score? An Example of Features

Different word
Same role
+5

Twist **facts** to suit **theories** , — *dobj*
NOUN *NOUN*

Same PoS-Tag
+10

theories to suit **facts.**
NOUN *NOUN*

Same word
Same role
-5

I weighted them with the — nsubj

only **weight I** could find.

How our algorithm sorts criss-cross patterns:
3 other representative examples of our 22 features

Problem 

Before 2015 there was no data to fit the system.

Problem

But the hand tuned systems of 2015+2016 allowed selective annotation: we have more annotated data than before! 3000/2M instances, with up to 31 Real Positives!

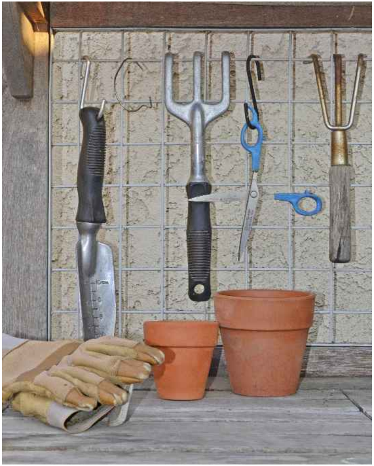
Is 0,15% of the corpus with only 31 true instances really enough to tune the weights automatically?

Problem 

But the hand tuned systems of 2015+2016 allowed selective annotation: we have more annotated data than before! 3000/2M instances, with up to 31 Real Positives!

Is 0,15% of the corpus with only 31 true instances really enough to tune the weights automatically?

Experimental Set Up



Experimental Set Up 

- Corpus
  - Parliament proceedings 
  - Training: 4 M words, 2 M instances, 3000 single annotated, 296 Doubly annotated, 31 Pos.
  - Test:2 M words, 1M instances
- Annotation
  - 2 annotators 
  - Instances true for both annotators=True.
  - Other instances (incl. unknown)=False
- Techniques and Tools
  - Evaluation with average precision
  - Stanford Parser and Tagger (CoreNLP)
  - Sklearn: Logistic regression

Results

Results 

Model		Avg Precision	Precision	Recall	F1-score
Machine	Base	57.1	80.0	30.8	44.4
Machine	All features	**70.8**	**90.0**	**69.2**	**78.3**
Human	Base	42.5	–	–	–
Human	All features	67.7	–	–	–

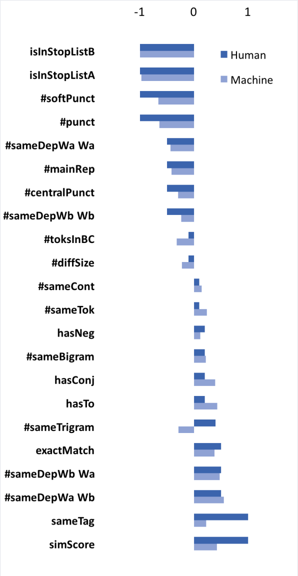
Results for logistic regression model (Machine) with comparison to the hand-tuned models of Dubremetz and Nivre (2015; 2016) (Human).

Evaluation with 13 Pos. instances in the test set.

Inter annotator agreement $\kappa = 0.69$

A system very precise with only borderline cases as false positives.

Results



-1
0
1
isInStopListB
isInStopListA
#softPunct
#punct
#sameDepWa Wa
#mainRep
#centralPunct
#sameDepWb Wb
#toksInBC
#diffSize
#sameCont
#sameTok
hasNeg
#sameBigram
hasConj
hasTo
#sameTrigram
exactMatch
#sameDepWb Wa
#sameDepWa Wb
sameTag
simScore
Human
Machine

Discussion and Perspectives

Future Work 

- Apply our method to other devices? Anaphora? Epiphora?
- Apply on other corpus

Contributions 

- Proof of concept: Machine Learning on Chiasmus is possible with little (but well chosen) data
- Additional knowledge about the features. Humans and Machine globally agree on what are the positive/negative features.

and

- Quotes, 36 000 quotes, 800 000 words
- Water Stone, 192 000 titles, 900 000 words (Literature corpus)
- DBLP, 192 000 titles, 2 M words (Computer science corpus)

Contributions 

- Proof of concept: Machine Learning on Chiasmus is possible with little (but well chosen) data
- Additional knowledge about the features. Humans and Machine globally agree on what are the positive/negative features.

and

- Quotes, 36 000 quotes, 800 000 words
- Water Stone, 192 000 titles, 900 000 words (Literature corpus)
- DBLP, 192 000 titles, 2 M words (Computer science corpus)

Take home message

Summary

Today you discovered:

- Chiasmus
- You don't need to annotate millions of examples to get a descent result for machine learning.
- Condition: preselect well your annotation pool through hand tuned features
- Promising results for detection of repetitive figures in general: feasible task at a low cost.

Thank You!

Questions?

Bonus: a Chiasmus Viewed by a Computer

If the mountain won't come to Mohammed, then let's take Mohammed to the mountain. **(In binary.)**

01001001 01100110 00100000 01110100 01101000 01100101 00100000 01101101 01101111 01110101 01101110 01110100 01100001 01101001 01101110 00100000 01110111 01101111 01101110 00100111 01110100 00100000 01100011 01101111 01101101 01100101 00100000 01110100 01101111 00100000 01001101 01101111 01101000 01100001 01101101 01101101 01100101 01100100 00101100 00100000 01110100 01101000 01100101 01101110 00100000 01101100 01100101 01110100 00100111 01110011 00100000 01110100 01100001 01101011 01100101 00100000 01001101 01101111 01101000 01100001 01101101 01101101 01100101 01100100 00100000 01110100 01101111 00100000 01110100 01101000 01100101 00100000 01101101 01101111 01110101 01101110 01110100 01100001 01101001 01101110 00101110

References

Dubremetz, M. & Nivre, J. (2015). Rhetorical Figure Detection: the Case of Chiasmus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, (pp. 23–31)., Denver, Colorado, USA. Association for Computational Linguistics.

Gawryjolek, J. J. (2009). *Automated Annotation and Visualization of Rhetorical Figures*. Master thesis, Universty of Waterloo.

Hromada, D. D. (2011). Initial Experiments with Multilingual Extraction of Rhetoric Figures by means of PERL-compatible Regular Expressions. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, (pp. 85–90)., Hissar, Bulgaria.

System, corpus, annotation available at: http://stp.lingfil.uu.se/~marie/chiasme.htm.

References

System, corpus, annotation available at:
http://stp.lingfil.uu.se/~marie/chiasme.htm.