# The proper place of men and machines in language technology

## Processing Russian without any linguistic knowledge

Serge Sharoff and Joakim Nivre

University of Leeds, University of Uppsala
Part financed by FP-7, STREP-248.005, and
LLP-KA2, 505630-LLP-1-2009-1-SE-KA2-KA2MP

27 May 2011

# Outline

# Corpora

## The procedure for all experiments

| Disambiguated RNC | | | SynTagRus | | |
|---|---|---|---|---|---|
| Tokens | Orth words | Sentences | Tokens | Orth words | Sentences |
| 5801316 | 5115016 | 432611 | 719957 | 635524 | 41186 |

# Corpora

## The procedure for all experiments

1. take an existing annotated corpus (RNC)

| Disambiguated RNC | | | SynTagRus | | |
|---|---|---|---|---|---|
| Tokens | Orth words | Sentences | Tokens | Orth words | Sentences |
| 5801316 | 5115016 | 432611 | 719957 | 635524 | 41186 |

# Corpora

## The procedure for all experiments

1. take an existing annotated corpus (RNC)
2. design a simplified representation

| Disambiguated RNC | | | SynTagRus | | |
|---|---|---|---|---|---|
| Tokens | Orth words | Sentences | Tokens | Orth words | Sentences |
| 5801316 | 5115016 | 432611 | 719957 | 635524 | 41186 |

# Corpora

## The procedure for all experiments

1. take an existing annotated corpus (RNC)
2. design a simplified representation
3. learn a model (in several iterations)
   Training, testing (and development) sets

| Disambiguated RNC | | | SynTagRus | | |
|---|---|---|---|---|---|
| Tokens | Orth words | Sentences | Tokens | Orth words | Sentences |
| 5801316 | 5115016 | 432611 | 719957 | 635524 | 41186 |

# Corpora

## The procedure for all experiments

1. take an existing annotated corpus (RNC)
2. design a simplified representation
3. learn a model (in several iterations)
   Training, testing (and development) sets

| Disambiguated RNC | | | SynTagRus | | |
|---|---|---|---|---|---|
| Tokens | Orth words | Sentences | Tokens | Orth words | Sentences |
| 5801316 | 5115016 | 432611 | 719957 | 635524 | 41186 |

- ruWac, produced by crawling Russian websites, and RNC
  parsed ruWac: 1797706747 tokens, 1254327 documents
  parsed RNC: 109326064 tokens, 35356 documents

# Outline

# Trigram model

- Morphology vs tags:
  шлепают: шлепать=V,несов=непрош,мн,изъяв,3-л
  шлепают: шлепать=Vmip3p-a-e
  4,592 feature combinations in the dRNC vs 956 tags

# Trigram model

- Morphology vs tags:
  шлепают: шлепать=V,несов=непрош,мн,изъяв,3-л
  шлепают: шлепать=Vmip3p-a-e
  4,592 feature combinations in the dRNC vs 956 tags
- Multext tagset (`http://nl.ijs.si/ME/V4/`)

# Trigram model

- Morphology vs tags:
  шлепают: шлепать=V,несов=непрош,мн,изъяв,3-л
  шлепают: шлепать=Vmip3p-a-e
  4,592 feature combinations in the dRNC vs 956 tags
- Multext tagset (`http://nl.ijs.si/ME/V4/`)
- Emission probabilities,
  *гравюра*: Ncfsnn 6
  *на*: Q 27 Sp-a 34821 Sp-l 24992
  *стали*: Ncfpan 1; Ncfpnn 1; Ncfsgn 23; Ncfsln 1; Vmis-p-a-e 1143

# Trigram model

- Morphology vs tags:
  шлепают: шлепать=V,несов=непрош,мн,изъяв,3-л
  шлепают: шлепать=Vmip3p-a-e
  4,592 feature combinations in the dRNC vs 956 tags
- Multext tagset (`http://nl.ijs.si/ME/V4/`)
- Emission probabilities,
  *гравюра*: Ncfsnn 6
  *на*: Q 27 Sp-a 34821 Sp-l 24992
  *стали*: Ncfpan 1; Ncfpnn 1; Ncfsgn 23; Ncfsln 1; Vmis-p-a-e 1143
- Transition probabilities

  `Ncfsnn Sp-l Ncfsln` 503;
  `Ncfsnn Sp-a Ncfpan` 23;
  `Ncfsnn Sp-l Vmi3ps-a-p` 0;
  `Ncfsnn Q Ncfsln` 0;

# Classes of tagging errors

Overall accuracy on full tags: 95.28% (4.72% error rate)

| Code | Explanation | Error rate | Relative error | Coverage |
|---|---|---|---|---|
| N | Nouns | 2.08% | 7.21% | 28.80% |
| A | Adjectives | 0.86% | 9.05% | 9.51% |
| P | Pronouns | 0.65% | 7.82% | 8.28% |
| V | Verbs | 0.50% | 4.89% | 10.16% |
| C | Conjunctions | 0.14% | 2.37% | 5.84% |
| R | Adverbs | 0.13% | 4.69% | 2.81% |
| S | Prepositions | 0.13% | 0.89% | 14.62% |
| M | Numerals | 0.13% | 4.60% | 2.81% |
| Q | Particles | 0.10% | 4.03% | 2.59% |
| I | Interjections | 0.01% | 26.42% | 0.02% |

# Most common incorrectly tagged words

| | | | |
|---|---|---|---|
| 0.0932% | TnT | как | C |
| 0.0920% | RNC | как | P-----r |
| 0.0788% | TnT | что | C |
| 0.0682% | TnT | ЭВМ | Ncfsgn |
| 0.0682% | RNC | ЭВМ | Ncfpgn |
| 0.0507% | RNC | что | P--nsnn |
| 0.0444% | TnT | это | P--nsnn |
| 0.0438% | TnT | как | P-----r |
| 0.0413% | TnT | судов | Ncnpgn |
| 0.0413% | RNC | судов | Ncmpgn |
| 0.0413% | RNC | как | C |
| 0.0363% | TnT | все | P--nsnn |
| 0.0357% | RNC | это | Q |
| 0.0350% | RNC | все | R |
| 0.0338% | RNC | что | P--nsan |
| 0.0325% | TnT | его | P-3msan |
| 0.0244% | RNC | что | C |
| 0.0244% | RNC | лиц | Ncnpgy |
| 0.0244% | TnT | лиц | Ncnpgn |
| 0.0238% | TnT | что | P--nsnn |

# Outline

# Learning lemmatisation rules

- Triples from dRNC: `tag, lemma, form`

# Learning lemmatisation rules

- Triples from dRNC: `tag, lemma, form`
- Finding the longest shared part for each pair **близ**кий-по**ближ**е; Rule *зкий←по*же

# Learning lemmatisation rules

- Triples from dRNC: `tag, lemma, form`
- Finding the longest shared part for each pair
  **близ**кий-по**ближ**е; Rule *зкий←по*же

ok пониже (*зкий←по*же) низкий.

# Learning lemmatisation rules

- Triples from dRNC: `tag, lemma, form`
- Finding the longest shared part for each pair **бли**зкий-по**бли**же; Rule *зкий←по*же

ok пониже (*зкий←по*же) низкий.

no похуже (*зкий←по*же) хузкий

# Learning lemmatisation rules

- Triples from dRNC: `tag, lemma, form`
- Finding the longest shared part for each pair **бли**зкий-по**бли**же; Rule *зкий←по*же

ok пониже (*зкий←по*же) низкий.

no похуже (*зкий←по*же) хузкий

- Frequency-based decision on more general rules

# Lemmatisation for Ncmsgy ending in -ц

| | | |
|---|---|---|
| ец | еца | кузнеца |
| иц | ица | фрица |
| заяц | зайца | |
| ец | йца | европейца |
| я-муромец | и-муромца | |
| ринц | ринца | принца |
| ртц | ртца | артца |
| ец | ьца | владельца |
| ец | ца | чеченца |

# Outline

# Training

- MaltParser: Dependency-based transition parser
  keeping stack for partially processed tokens and adding arcs

# Training

- MaltParser: Dependency-based transition parser
  keeping stack for partially processed tokens and adding arcs
- Machine Learning for actions:
  linear vs polynomial SVM

# Training

- MaltParser: Dependency-based transition parser
  keeping stack for partially processed tokens and adding arcs
- Machine Learning for actions:
  linear vs polynomial SVM
- Five attributes: POS, dep, mor, lem, lex

# Training

- MaltParser: Dependency-based transition parser
  keeping stack for partially processed tokens and adding arcs
- Machine Learning for actions:
  linear vs polynomial SVM
- Five attributes: POS, dep, mor, lem, lex
- Integration with the tagger:
  Using MTE tags
  Disregarding multiword expressions and phantom nodes

# Assessing the accuracy on SynTagRus

- Unlabelled attachment score (UAS) vs Labelled attachment score (LAS)

| | **LAS** | **UAS** |
|---|---|---|
| SynTagRus tags, poly-SVM | 83.4 | 89.4 |
| MTE tags, poly-SVM | 82.8 | 88.8 |
| MTE tags, linear SVM | 82.2 | 88.0 |

# Parsing example (on RNC)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | " | " | - | - | - | 0 | PUNC | ROOT | |
| 2 | Ты | ты | P | P | P-2-snn | 4 | предик | предик | |
| 3 | не | не | Q | Q | Q | 4 | огранич | огранич | |
| 4 | поверишь | | поверить | | V | V | Vmif2s-a-p | 0 | ROO |
| 5 | , | , | , | , | , | 4 | PUNC | PUNC | |
| 6 | как | как | C | C | C | 8 | обст | вводн | |
| 7 | мне | я | P | P | P-1-sdn | 8 | дат-субъект | дат-субъект | |
| 8 | грустно | грустно | R | R | R | 4 | 1-компл | сравн-союзн | |
| 9 | , | , | , | , | , | 8 | PUNC | PUNC | |
| 10 | что | что | C | C | C | 8 | предик | предик | |
| 11 | мой | мой | P | P | P--msna | 13 | опред | опред | |
| 12 | добрый | добрый | A | A | Afpmsnf | 13 | опред | опред | |
| 13 | Василий | василий | N | N | Npmsny | 15 | предик | предик | |
| 14 | Дмитриевич | | дмитриевич | | N | N | Npmsny | 13 | аппоз | апп |
| 15 | уехал | уехать | V | V | Vmis-sma-p | | 10 | подч-союзн | под |
| 16 | вот | вот | Q | Q | Q | 18 | огранич | огранич | |
| 17 | уже | уже | R | R | R | 18 | огранич | огранич | |
| 18 | неделя | неделя | N | N | Ncfsnn | 15 | обст | обст | |
| 19 | . | . | S | S | SENT | 18 | PUNC | PUNC | |

# Parsing example (on RNC)

```
1       Какую     какой     P       P       P--fsaa 2       опред
2       поэзию    поэзия    N       N       Ncfsan  3       1-компл
3       любит     любить    V       V       Vmip3s-a-e      0       ROOT
4       Андропов            андропов        N       N       Npmsny  3       предик
5       ?         ?         S       S       SENT    4       PUNC

1       Пушкина   пушкин    N       N       Npmsgy  0       ROOT
2       за        за        S       S       Sp-l    0       ROOT
3       его       его       P       P       P-----a 4       квазиагент
4       слова     слово     N       N       Ncnpan  2       предл
5       :         :         -       -       -       4       PUNC
6       «         «         -       -       -       0       ROOT
7       Души      душа      N       N       Ncfpnn  6       предл
8       прекрасные          прекрасный      A       A       Afpmpnf 9       опред
9       порывы    порыв     N       N       Ncmpnn  6       сочин
10      !         !         S       S       SENT    9       PUNC
```

# Most common errors in SynTagRus

| | | |
|---|---|---|
| 0.403% | PUNC | ROOT |
| 0.148% | 1-компл | обст |
| 0.142% | 2-компл | обст |
| 0.133% | 1-компл | квазиагент |
| 0.116% | обст | 1-компл |
| 0.113% | 2-компл | 1-компл |
| 0.105% | предик | 1-компл |
| 0.105% | квазиагент | 1-компл |
| 0.100% | сочин | сент-соч |
| 0.096% | атриб | обст |
| 0.080% | обст | атриб |
| 0.077% | предик | ROOT |
| 0.077% | атриб | квазиагент |
| 0.074% | обст | ROOT |
| 0.074% | 1-компл | предик |
| 0.073% | обст | огранич |
| 0.068% | сочин | обст |
| 0.068% | атриб | 1-компл |

# Genre-specific relations in RNC: nonfiction

| Relation | Total RNC | Nonfiction | LL-score |
|---|---|---|---|
| квазиагент | 3066600 | 879725 | 291893 |
| опред | 9970254 | 2086966 | 143706 |
| нум-аппоз | 162837 | 90701 | 130746 |
| примыкат | 301102 | 131814 | 125249 |
| кратн | 140789 | 80072 | 119274 |
| количест | 847179 | 231689 | 64995 |
| атриб | 1934951 | 408153 | 29824 |
| аппоз | 2744305 | 547575 | 24822 |
| агент | 180813 | 54375 | 20998 |
| компл-аппоз | 31171 | 12573 | 10150 |
| пасс-анал | 176427 | 37894 | 3161 |
| композ | 37969 | 9212 | 1562 |
| ном-аппоз | 11570 | 2912 | 585 |
| распред | 28409 | 6108 | 513 |
| изъясн | 37747 | 7441 | 299 |
| 4-компл | 9841 | 2114 | 177 |

# Summary and resources

- Reasonably accurate taggers, lemmatisers and parsers
- Reasonably fast and reliable
- Available from `http://corpus.leeds.ac.uk/tools/ru/`
- Parsed ruWac to be available soon as well

# Was it really without linguistic knowledge?

- Flective languages
  agglutinative (data sparseness) or isolating (greater ambiguity and harder guessing);

# Was it really without linguistic knowledge?

- Flective languages
  agglutinative (data sparseness) or isolating (greater ambiguity and harder guessing);
- Unknown words: *vociferation, votazione, конъюгация*, 自由主义

# Was it really without linguistic knowledge?

- Flective languages
  agglutinative (data sparseness) or isolating (greater ambiguity and harder guessing);
- Unknown words: *vociferation*, *votazione*, *конъюгация*, 自由主义
- Generalisation over tag labels: `Afpmsg` `Ncmsgn/Ncmsgy/Npmsgy` `Afpfsd` `Ncfsdn/Ncmsdy/Npmsdy`

# Was it really without linguistic knowledge?

- Flective languages
  agglutinative (data sparseness) or isolating (greater ambiguity and harder guessing);
- Unknown words: *vociferation, votazione, конъюгация*, 自由主义
- Generalisation over tag labels: `Afpmsg` `Ncmsgn/Ncmsgy/Npmsgy` `Afpfsd` `Ncfsdn/Ncmsdy/Npmsdy`
- Generalisation over genre patterns: 98% on WSJ vs 85.7% on Internet forums (Giesbrecht, Evert, 2009),

# Was it really without linguistic knowledge?

- Flective languages
  agglutinative (data sparseness) or isolating (greater ambiguity and harder guessing);
- Unknown words: *vociferation, votazione, конъюгация*, 自由主义
- Generalisation over tag labels: `Afpmsg` `Ncmsgn/Ncmsgy/Npmsgy` `Afpfsd` `Ncfsdn/Ncmsdy/Npmsdy`
- Generalisation over genre patterns: 98% on WSJ vs 85.7% on Internet forums (Giesbrecht, Evert, 2009),

  Possible solution with domain adaptation

# Was it really without linguistic knowledge?

- Flective languages
  agglutinative (data sparseness) or isolating (greater ambiguity and harder guessing);
- Unknown words: *vociferation, votazione, конъюгация*, 自由主义
- Generalisation over tag labels: `Afpmsg` `Ncmsgn/Ncmsgy/Npmsgy` `Afpfsd` `Ncfsdn/Ncmsdy/Npmsdy`
- Generalisation over genre patterns: 98% on WSJ vs 85.7% on Internet forums (Giesbrecht, Evert, 2009),

  Possible solution with domain adaptation
- Accuracy on cross-validation is not the only answer

# Was it really without linguistic knowledge?

- Flective languages
  agglutinative (data sparseness) or isolating (greater ambiguity and harder guessing);
- Unknown words: *vociferation, votazione, конъюгация*, 自由主义
- Generalisation over tag labels: `Afpmsg` `Ncmsgn/Ncmsgy/Npmsgy` `Afpfsd` `Ncfsdn/Ncmsdy/Npmsdy`
- Generalisation over genre patterns: 98% on WSJ vs 85.7% on Internet forums (Giesbrecht, Evert, 2009),

  Possible solution with domain adaptation
- Accuracy on cross-validation is not the only answer
- Correction of errors: no rules