Capitalization Cues Improve Dpendency Grammar Induction

Valentin I. Spitkovsky

with **Daniel Jurafsky** (Stanford University)
and **Hiyan Alshawi** (Google Inc.)

Problem: Grammar Induction is Hard

Problem: Grammar Induction is Hard

Major challenges:

Problem: Grammar Induction is Hard

Major challenges:

- **non-convex objectives**

Problem: Grammar Induction is Hard

Major challenges:

- **non-convex objectives**
- **poor correlations between likelihood and accuracy**

Problem: Grammar Induction is Hard

Major challenges:

- **non-convex objectives** **(Gimpel and Smith, 2012)**
- **poor correlations between likelihood and accuracy**

Problem: Grammar Induction is Hard

Major challenges:

- **non-convex objectives** (Gimpel and Smith, 2012)
- **poor correlations between likelihood and accuracy** (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)

Problem: Grammar Induction is Hard

Major challenges:

- **non-convex objectives** **(Gimpel and Smith, 2012)**
- **poor correlations between likelihood and accuracy**
  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)
  - **e.g., optimizers run away from supervised MLE solutions**

Problem: Grammar Induction is Hard

Major challenges:

- **non-convex objectives** **(Gimpel and Smith, 2012)**
- **poor correlations between likelihood and accuracy**
  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)
  - **e.g., optimizers run away from supervised MLE solutions**
    (to the tune of 20 points of accuracy)

Problem: Grammar Induction is Hard

Major challenges:

- **non-convex objectives** **(Gimpel and Smith, 2012)**
- **poor correlations between likelihood and accuracy** **(Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)**
  - **e.g., optimizers run away from supervised MLE solutions (to the tune of 20 points of accuracy)**
- **flaws in evaluation** **(Schwartz et al., 2011)**

Problem: Grammar Induction is Hard

Major challenges:

- **non-convex objectives** (Gimpel and Smith, 2012)
- **poor correlations between likelihood and accuracy** (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)
  - e.g., optimizers run away from supervised MLE solutions (to the tune of 20 points of accuracy)
- **flaws in evaluation** (Schwartz et al., 2011)

Partial solutions:

Problem: Grammar Induction is Hard

Major challenges:

- non-convex objectives (Gimpel and Smith, 2012)
- poor correlations between likelihood and accuracy (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)
  - e.g., optimizers run away from supervised MLE solutions (to the tune of 20 points of accuracy)
- flaws in evaluation (Schwartz et al., 2011)

Partial solutions:

- train on more / better data (Mareček and Zabokrtský, 2012)

Problem: Grammar Induction is Hard

Major challenges:

- non-convex objectives (Gimpel and Smith, 2012)
- poor correlations between likelihood and accuracy (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)
  - e.g., optimizers run away from supervised MLE solutions (to the tune of 20 points of accuracy)
- flaws in evaluation (Schwartz et al., 2011)

Partial solutions:

- train on more / better data (Mareček and Zabokrtský, 2012)
- test many data sets / languages (fight noise with CLT)

Problem: Grammar Induction is Hard

Major challenges:

- non-convex objectives (Gimpel and Smith, 2012)
- poor correlations between likelihood and accuracy (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)
  - e.g., optimizers run away from supervised MLE solutions (to the tune of 20 points of accuracy)
- flaws in evaluation (Schwartz et al., 2011)

Partial solutions:

- train on more / better data (Mareček and Zabokrtský, 2012)
- test many data sets / languages (fight noise with CLT)
- employ less ad-hoc initializers ("eat your own dog food")

Problem: Grammar Induction is Hard

Major challenges:

- non-convex objectives (Gimpel and Smith, 2012)
- poor correlations between likelihood and accuracy (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994; Liang and Klein, 2008; Spitkovsky et al., 2009–2011)
  - e.g., optimizers run away from supervised MLE solutions (to the tune of 20 points of accuracy)
- flaws in evaluation (Schwartz et al., 2011)

Partial solutions:

- train on more / better data (Mareček and Zabokrtský, 2012)
- test many data sets / languages (fight noise with CLT)
- employ less ad-hoc initializers ("eat your own dog food")
- constrain search space (structure is underdetermined)

Idea: Use Capitalization as Parsing Cues

Idea: Use Capitalization as Parsing Cues

Partial bracketing constraints: **(Pereira and Schabes, 1992)**

Idea: Use Capitalization as Parsing Cues

Partial bracketing constraints: (Pereira and Schabes, 1992)

- **semantic annotations** (Naseem and Barzilay, 2011)
- **punctuation marks** (Ponvert et al., 2010)
- **web markup** (Spitkovsky et al., 2010)

Idea: Use Capitalization as Parsing Cues

Partial bracketing constraints: **(Pereira and Schabes, 1992)**

- **semantic annotations** **(Naseem and Barzilay, 2011)**
- **punctuation marks** **(Ponvert et al., 2010)**
- **web markup** **(Spitkovsky et al., 2010)**

... defined over raw text (no POS tags).

Example: (no punctuation, etc. cues)

Example:

(no punctuation, etc. cues)

[$_{NP}$ Jay Stevens] of [$_{NP}$ Dean Witter] actually cut his per-share earnings estimate to [$_{NP}$ \$9] from [$_{NP}$ \$9.50] for [$_{NP}$ 1989] and to [$_{NP}$ \$9.50] from [$_{NP}$ \$10.35] in [$_{NP}$ 1990] because he decided sales would be even weaker than he had expected.

Example: (less WSJ-ish)

Example:

(less WSJ-ish)

$[_{NP}$ Jurors$]$ in $[_{NP}$ U.S. District Court$]$ in $[_{NP}$ Miami$]$ cleared $[_{NP}$ Harold Hershhenson$]$, a former executive vice president; $[_{NP}$ John Pagones$]$, a former vice president; and $[_{NP}$ Stephen Vadas$]$ and $[_{NP}$ Dean Ciporkin$]$, who had been engineers with $[_{NP}$ Cordis$]$.

Analysis: (English PTB)

- **Mostly noun phrases (96%):**

Analysis: (English PTB)

- **Mostly noun phrases (96%):**

  Apple II
  World War I
  Mayor William H. Hudnut III
  International Business Machines Corp.
  Alexandria, Va

Analysis: (English PTB)

- **Mostly noun phrases (96%):**

  Apple II
  World War I
  Mayor William H. Hudnut III
  International Business Machines Corp.
  Alexandria, Va

- **Some proper adjectives (5%);**

Analysis: (English PTB)

- **Mostly noun phrases (96%):**

  Apple II
  World War I
  Mayor William H. Hudnut III
  International Business Machines Corp.
  Alexandria, Va

- **Some proper adjectives (5%);**

- **First-person pronoun, I (2%).**

Analysis: (English PTB)

- **Mostly noun phrases (96%):**

  Apple II
  World War I
  Mayor William H. Hudnut III
  International Business Machines Corp.
  Alexandria, Va

- **Some proper adjectives (5%);**

- **First-person pronoun, I (2%).**

— Yields more accurate dependency parsing constraints than either markup or punctuation (for WSJ).

Experiments: (CoNLL 2006/7)

- **Data:**

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**
  - **not Spanish or Basque (because of post-processing)**

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**
  - **not Spanish or Basque (because of post-processing)**
  - **not Japanese, Chinese or Arabic...**

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**
  - **not Spanish or Basque (because of post-processing)**
  - **not Japanese, Chinese or Arabic...**

- **Model:**

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**
  - **not Spanish or Basque (because of post-processing)**
  - **not Japanese, Chinese or Arabic...**

- **Model:**
  - **DBM-1** **(Spitkovsky et al., 2012)**

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**
  - **not Spanish or Basque (because of post-processing)**
  - **not Japanese, Chinese or Arabic...**

- **Model:**
  - **DBM-1 (Spitkovsky et al., 2012)**
  - **first dependency-and-boundary model (see EMNLP)**

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**
  - **not Spanish or Basque (because of post-processing)**
  - **not Japanese, Chinese or Arabic...**

- **Model:**
  - **DBM-1 (Spitkovsky et al., 2012)**
  - **first dependency-and-boundary model (see EMNLP)**

- **Training:**

Experiments: (CoNLL 2006/7)

- **Data:**
  - 14 languages with case information
  - not Spanish or Basque (because of post-processing)
  - not Japanese, Chinese or Arabic...
- **Model:**
  - DBM-1 (Spitkovsky et al., 2012)
  - first dependency-and-boundary model (see EMNLP)
- **Training:**
  - vanilla EM

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**
  - **not Spanish or Basque (because of post-processing)**
  - **not Japanese, Chinese or Arabic...**

- **Model:**
  - **DBM-1** (Spitkovsky et al., 2012)
  - **first dependency-and-boundary model** (see EMNLP)

- **Training:**
  - **vanilla EM**
  - **controls: uniform Viterbi init** (Cohen and Smith, 2010)

Experiments: (CoNLL 2006/7)

- **Data:**
  - **14 languages with case information**
  - **not Spanish or Basque (because of post-processing)**
  - **not Japanese, Chinese or Arabic...**
- **Model:**
  - **DBM-1** (Spitkovsky et al., 2012)
  - **first dependency-and-boundary model** (see EMNLP)
- **Training:**
  - **vanilla EM**
  - **controls: uniform Viterbi init** (Cohen and Smith, 2010)
  - **capitalization: constrained sampling of initial parse trees**

Results:

Results:

- 2^+ increase in accuracy

Results:

- 2^+ **increase in accuracy (on average, 42.8 $\rightarrow$ 45)**

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline
  - with various different constraints

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline
  - with various different constraints
  - helps in training and during inference

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline
  - with various different constraints
  - helps in training and during inference
  - and also in combination with punctuation

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline
  - with various different constraints
  - helps in training and during inference
  - and also in combination with punctuation

- but, most of the gain is from just two languages...

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline
  - with various different constraints
  - helps in training and during inference
  - and also in combination with punctuation

- **but**, most of the gain is from just two languages...
  - Italian (+11) and Greek (+18)

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline
  - with various different constraints
  - helps in training and during inference
  - and also in combination with punctuation

- **but**, most of the gain is from just two languages...
  - Italian (+11) and Greek (+18)
  - worst impact on English (-0.02)

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline
  - with various different constraints
  - helps in training and during inference
  - and also in combination with punctuation

- but, most of the gain is from just two languages...
  - Italian (+11) and Greek (+18)
  - worst impact on English (-0.02), so much for inspiration...

Results:

- 2^+ increase in accuracy (on average, $42.8 \rightarrow 45$)
  - over a state-of-the-art baseline
  - with various different constraints
  - helps in training and during inference
  - and also in combination with punctuation

- but, most of the gain is from just two languages...
  - Italian (+11) and Greek (+18)
  - worst impact on English (-0.02), so much for inspiration...
  - still, virtually no harm — even in the worst case!

Conclusion:

Conclusion:

- **informative signal, but requires further investigation**

Conclusion:

- **informative signal, but requires further investigation**
  - **very preliminary results...**

Conclusion:

- **informative signal, but requires further investigation**
  - **very preliminary results...**
  - **cues may be more useful as features!**

Conclusion:

- informative signal, but requires further investigation
  - very preliminary results...
  - cues may be more useful as features!

- miscellaneous observations:

Conclusion:

- **informative signal, but requires further investigation**
  - **very preliminary results...**
  - **cues may be more useful as features!**

- **miscellaneous observations:**
  - **transitions between scripts**

Conclusion:

- **informative signal, but requires further investigation**
  - **very preliminary results...**
  - **cues may be more useful as features!**

- **miscellaneous observations:**
  - **transitions between scripts**
    - **e.g., for Arabic, CJK, numerals, etc.**

Conclusion:

- informative signal, but requires further investigation
  - very preliminary results...
  - cues may be more useful as features!

- miscellaneous observations:
  - transitions between scripts
    - e.g., for Arabic, CJK, numerals, etc.
  - interaction with punctuation / “operator” precedence

Conclusion:

- **informative signal, but requires further investigation**
  - **very preliminary results...**
  - **cues may be more useful as features!**

- **miscellaneous observations:**
  - **transitions between scripts**
    - **e.g., for Arabic, CJK, numerals, etc.**

  - **interaction with punctuation / “operator” precedence**
    - **e.g., Alexandria, Va**

Conclusion:

- informative signal, but requires further investigation
  - very preliminary results...
  - cues may be more useful as features!

- miscellaneous observations:
  - transitions between scripts
    - e.g., for Arabic, CJK, numerals, etc.
  - interaction with punctuation / “operator” precedence
    - e.g., Alexandria, Va
      vs. Kawasaki Heavy Industries Ltd.,
      Mitsubishi Heavy Industries Ltd. and ...

Conclusion:

- informative signal, but requires further investigation
  - very preliminary results...
  - cues may be more useful as features!

- miscellaneous observations:
  - transitions between scripts
    - e.g., for Arabic, CJK, numerals, etc.
  - interaction with punctuation / “operator” precedence
    - e.g., Alexandria, Va
      vs. Kawasaki Heavy Industries Ltd.,
      Mitsubishi Heavy Industries Ltd. and ...
  - properties of first (and last) words

Thanks!

No questions at this time...