The Language Resource Spectrum: A Perspective from Google

Ryan McDonald

Google NLU team
Google Linguistics team
The Language Resource Spectrum:
A Perspective from Google

Ryan McDonald

Google NLU team
Google Linguistics team
Language Resource Spectrum
Language Resource Spectrum

- World Wide Web
- Wikipedia (The Free Encyclopedia)
- Wiktionary (The Free Dictionary)
- Search logs
- Yelp
- European Parliament
- Fully supervised
- Weakly supervised
- Unsupervised
amount of data

supervision
amount of data vs. supervision
amount of data

which is better?

supervision
High quality annotations
Crowd-sourced
Pre-existing resources
Software Engineer
Auto resources
Models
Active Learning
High quality annotations  Crowd-sourced  Pre-existing resources  Software Engineer
Auto resources  Models  Active Learning
High quality annotations

Crowd-sourced

Pre-existing resources

Software Engineer
Auto resources
Models
Active Learning
ML is really good at the head
Pipelined / multi-Component Systems

- Indexer
- Text Extractor
- Segmentation
- Morphosyntax
- Shallow Semantics
- Entity resolution
- Relation extraction
- MT
- QA
- Search
Pipelined / multi-Component Systems

Indexer
Text Extractor

Segmentation
Morphosyntax

Shallow Semantics

Entity resolution
Relation extraction

End User Task
MT
QA
Search
Pipelined / multi-Component Systems

Indexer
Text Extractor
Segmentation
Morphosyntax

Shallow Semantics
Entity resolution
Relation extraction

MT
QA
Search

Upstream Task

End User Task

LREC, May 2016
Upstream: Morphosyntactactic Tagging
### Feature-based Classification

<table>
<thead>
<tr>
<th>ce</th>
<th>que</th>
<th>vous</th>
<th>n</th>
<th>entendrez</th>
<th>jamais</th>
<th>dans</th>
<th>un</th>
<th>débat</th>
<th>politique</th>
<th>français</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDEM PRON POS=PRON number=sing</td>
<td>IN ADP</td>
<td>PRP PRON POS=PRON number=plur person=2</td>
<td>RB ADV POS=ADV</td>
<td>entendre VBC VERB POS=VERB number=plur person=2 tense=fut mood=ind</td>
<td>RB ADV POS=ADV</td>
<td>IN ADP POS=ADP</td>
<td>DT DET POS=DET gender=masc number=sing</td>
<td>NN NOUN POS=NOUN gender=masc number=sing</td>
<td>JJ ADJ POS=ADJ gender=masc number=sing</td>
<td>JJ ADJ POS=ADJ gender=masc number=sing</td>
</tr>
</tbody>
</table>
Feature-based Classification

word=entendrez
suffix3=rez
word-1=n
word+1=jamais
cluster=124
cluster-1=53
cluster+1=210
Resource Trade-Off

Annotated data

Model
Resource Trade-Off

Annotated data → Model → Dictionaries / Lexicons
Morphosyntactic Lexicons via Graph-Propagation
Morphosyntactic Lexicons via Graph-Propagation
Morphosyntactic Lexicons via Graph-Propagation
Morphosyntactic Lexicons via Graph-Propagation
Morphosyntactic Lexicons via Graph-Propagation

Number=Plur

candies

suff:ies:y
pref

candy

clust:10

suff:s

dish

cats

clust:35

suff:s:
pref

cats

dishes

clust:20

suff:es:
pref

LREC, May 2016
Morphosyntactic Lexicons via Graph-Propagation

Number=Plur

candies
-1

candy
+1

cat

cats

dish
-1

dishes

Clust: 10
Suff: ies: y
Pref

Clust: 10
Suff: s

Clust: 20
Suff: es:
Pref

Clust: 35
Suff: s
Pref
Morphosyntactic Lexicons via Graph-Propagation

Number=Plur

Reinforce (1)
- suff:s
- clust:35
- clust:20

Flip (-1)
- suff:ies:y
- suff:es:
- suff:s:

Neutral (0)
- pref
- clust:10
Morphosyntactic Lexicons via Graph-Propagation

Number = Plur

Reinforce (1)
- suff:s
- clust:35
- clust:20

Flip (-1)
- suff:ies:y
- suff:es:
- suff:s:

Neutral (0)
- pref
- clust:10
Morphosyntactic Lexicons via Graph-Propagation

Number=Plur

Reinforce (1)
- suff:s
- clust:35
- clust:20

Flip (-1)
- suff:ies:y
- suff:es:
- suff:s:

Neutral (0)
- pref
- clust:10

Faruqui et al '16: Ising Mean Field Approximation

Morphosyntactic Lexicons via Graph-Propagation

LREC, May 2016
Universal Lexicons

* Seed with Universal Dependencies (Nivre et al. ’16)

```
John saw Mary
The saw broke
NOUN: Number=Sing
VERB: Tense=Past
saw
```
Resource Trade-off

- Baseline*
- +Auto Lexicon
- +Gold
- + Gold + Auto
- + 100% data

For each resource:
- Cs
  - Baseline: 95.5
  - +Auto Lexicon: 91.7
  - +Gold: 91
  - + Gold + Auto: 91
  - + 100% data: 91

- Fi
  - Baseline: 95.5
  - +Auto Lexicon: 91.7
  - +Gold: 91
  - + Gold + Auto: 91
  - + 100% data: 91

- Hu
  - Baseline: 95.5
  - +Auto Lexicon: 91.7
  - +Gold: 91
  - + Gold + Auto: 91
  - + 100% data: 91
**Resource Trade-off**

- Baseline*
- +Auto Lexicon
- +Gold
- + Gold + Auto
- + 100% data

<table>
<thead>
<tr>
<th></th>
<th>Cs</th>
<th>Fi</th>
<th>Hu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline*</td>
<td>96.8</td>
<td>93.6</td>
<td>95.5</td>
</tr>
<tr>
<td>+Auto Lexicon</td>
<td>97</td>
<td>95.2</td>
<td>95.1</td>
</tr>
<tr>
<td>+Gold</td>
<td>95.1</td>
<td>93.2</td>
<td>91</td>
</tr>
<tr>
<td>+ Gold + Auto</td>
<td>97</td>
<td>93.6</td>
<td>91</td>
</tr>
<tr>
<td>+ 100% data</td>
<td>97</td>
<td>95.2</td>
<td>93.2</td>
</tr>
</tbody>
</table>
Resource Trade-off

Baseline*  +Auto Lexicon  +Gold  + Gold + Auto  + 100% data

<table>
<thead>
<tr>
<th>Cs</th>
<th>Fi</th>
<th>Hu</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.5</td>
<td>91.7</td>
<td>91</td>
</tr>
<tr>
<td>96.8</td>
<td>93.6</td>
<td>93.2</td>
</tr>
<tr>
<td>97</td>
<td>95.2</td>
<td>95.1</td>
</tr>
<tr>
<td>97.4</td>
<td>95.7</td>
<td>95.8</td>
</tr>
</tbody>
</table>
### Resource Trade-off

- **Baseline**
- +Auto Lexicon
- +Gold
- + Gold + Auto
- + 100% data

<table>
<thead>
<tr>
<th>Resource</th>
<th>Baseline*</th>
<th>+Auto Lexicon</th>
<th>+Gold</th>
<th>+ Gold + Auto</th>
<th>+ 100% data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cs</td>
<td>96.8</td>
<td>97</td>
<td>97.4</td>
<td>96.8</td>
<td></td>
</tr>
<tr>
<td>Fi</td>
<td>95.5</td>
<td>93.6</td>
<td>96.8</td>
<td>95.2</td>
<td>95.7</td>
</tr>
<tr>
<td>Hu</td>
<td>97</td>
<td>92.8</td>
<td>93.2</td>
<td>91</td>
<td>95.8</td>
</tr>
</tbody>
</table>
Part-Of-Speech Tagging: Queries
Search Logs

Google search for "breaking bad" shows results including the AMC's official site for Breaking Bad, a Wikipedia entry, and a tweet from BreakingBad_AMC. The Wikipedia entry states that Breaking Bad is an American crime drama television series created and produced by Vince Gilligan. The AMC network aired it for five seasons. IMDb rates the show 9.5/10 with 847,508 votes. The final episode aired on September 29, 2013, and it is a spin-off of Better Call Saul. Awards include the Primetime Emmy Award for Outstanding Drama Series.
Search Logs

Google

breaking bad

Breaking Bad – AMC - AMC.com
www.amc.com/shows/breaking-bad
The official site for AMC's critically-acclaimed series Breaking Bad: Get full episodes, games, videos, plus episode & character guides.

Breaking Bad - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Breaking_Bad
Breaking Bad is an American crime drama television series created and produced by Vince Gilligan. The show originally aired on the AMC network for five ...
List of Breaking Bad episodes · Bryan Cranston · Breaking Bad (season 5) · RJ Mitte

Breaking Bad (@BreakingBad_AMC) | Twitter
https://twitter.com/BreakingBad_AMC

Breaking Bad (TV Series 2008–2013) - IMDb
www.imdb.com/title/tt0903747/

Breaking Bad -- Events set in motion long ago move toward a conclusion. Photos. Betsy Brandt at Breaking Bad (2008) Luis Moncada and Daniel Moncada in ...
Full Cast & Crew · 62 Episodes · Season 2 · Season 4

9.5/10 IMDb
95% Rotten Tomatoes

Breaking Bad
American drama series

Final episode date: September 29, 2013
Spin-off: Better Call Saul
Awards: Primetime Emmy Award for Outstanding Drama Series, More
POS Taggers From Clicks
Ganchev et al. (2012)

- MS-251-NVX: Baseline 92.8, +Click 81.9, Long Tail 74.3
POS Taggers From Clicks
Ganchev et al. (2012)

Baseline

+Click

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>+Click</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-251-NVX</td>
<td>92.8</td>
<td>93.5</td>
</tr>
<tr>
<td>MS-251</td>
<td>81.9</td>
<td>84.5</td>
</tr>
<tr>
<td>Long Tail</td>
<td>74.3</td>
<td>77.5</td>
</tr>
</tbody>
</table>
Morphosyntax Conclusions

- Money on more supervised data not necessarily optimal
- Better alternative: lexical resources (auto, manual & both)
- Better alternative: correlate usage statistics (click logs)
End User: Machine Translation
Pipelined Machine Translation

Preprocess → Reorder → Translate → Postorder → Postprocess

John saw Mary → John Mary saw

Preprocess

Reorder

Translate

ジョンはメアリーを見ました

LREC, May 2016
Pipelined Machine Translation

Preprocess → Reorder → Translate → Postorder → Postprocess

John saw Mary

Preprocess

ジョンはメアリーを見ました

Translate

John Mary saw

Reorder
Reordering Data vs.

Syntax-based reordered
(Lerner & Petrov ’13)

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>BLEU Δ</th>
<th>Human</th>
<th>Auto</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-&gt;Ar</td>
<td>1.2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>En-&gt;Iw</td>
<td>1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>En-&gt;Ja</td>
<td>3.5</td>
<td>3.7</td>
<td></td>
</tr>
</tbody>
</table>
Reordering Data VS.

Syntax-based reordered
(Lerner & Petrov ’13)

- **En->Ar**
  - Human: 1
  - Auto: 1.2
  - Rule-based: 1.5

- **En->Iw**
  - Human: 1.4
  - Auto: 1.4

- **En->Ja**
  - Human: 3.5
  - Rule-based: 3.7
  - Auto: 4.8

- **En->Ja**
  - Rule-based: 3.5
  - Auto: 3.7
**Better Parsers**

<table>
<thead>
<tr>
<th>Time</th>
<th>MT Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>74.5</td>
<td>77</td>
</tr>
<tr>
<td>77</td>
<td>79.5</td>
</tr>
<tr>
<td>79.5</td>
<td>82</td>
</tr>
</tbody>
</table>

- **Structured Training**
- **Greedy transition-based**
- **SSL**
- **Better features**
Better Parsers

**MT QUALITY**

- Greedy transition-based
- Structured Training
- SSL
- Better features

1pt improvement is significant to humans
More Data

Syntax

Reordering

EN->JA Reordering Score

Katz-Brown et al; Hall et al. 2011
Machine Translation

- Human vs. auto data: about the same
- Human models sometimes better than learned
- Better parsing models = better translation
- Better to **spend on targeted resources** — reordering
End User: Sentence Compression
Former Los Angeles Lakers head coach Phil Jackson won eleven NBA championships. He won six titles with the Chicago Bulls and five titles with the Lakers.
Google News

Filippova & Altun '13

- Can extract millions of pairs
- Quality ~ expert annotations
- 81.4 -> 84.3 F1 (10% -> 100% data)
Need For High Quality Annotations?
Need For High Quality Annotations?

Filippova et al. ’15: LSTM compression by deletion
**Need For High Quality Annotations?**

Filippova et al. ‘15: LSTM compression by deletion

Word [ ] [ ] [ ] [ ] [ ]

Accuracy 30

1/0
Need for High Quality Annotations?

Filippova et al. ’15: LSTM compression by deletion

Accuracy

LREC, May 2016
Need For High Quality Annotations?

Filippova et al. ’15: LSTM compression by deletion

Accuracy

30

31

34
The Resource Trade-off

High quality annotations → Crowd-sourced annotations → Auto resources

Data + model → Data + model
End User: QA & Knowledge Extraction
Jacob Bernoulli

Mathematician

Jacob Bernoulli was one of the many prominent mathematicians in the Bernoulli family. He was an early proponent of Leibnizian calculus and had sided with Leibniz during the Leibniz–Newton calculus controversy. Wikipedia

Born: January 6, 1655, Basel, Switzerland
Died: August 16, 1705, Old Swiss Confederacy
Education: University of Basel
Siblings: Johann Bernoulli
Doctoral students: Johann Bernoulli, Jakob Hermann, Nicolaus I Bernoulli

Nephews
Daniel Bernoulli
through Johann Bernoulli
Nicolaus II Bernoulli
through Johann Bernoulli
Johann II Bernoulli
through Johann Bernoulli
### QA & Knowledge Extraction

#### Jacob Bernoulli

**Mathematician**

Jacob Bernoulli was one of the many prominent mathematicians in the Bernoulli family. He was an early proponent of Leibnizian calculus and had sided with Leibniz during the Leibniz–Newton calculus controversy.

- **Born:** January 6, 1655, Basel, Switzerland
- **Died:** August 16, 1705, Old Swiss Confederacy
- **Education:** University of Basel
- **Siblings:** Johann Bernoulli
- **Doctoral students:** Johann Bernoulli, Jakob Hermann, Nicolaus Bernoulli
- **Nephews:**
  - Daniel Bernoulli through Johann Bernoulli
  - Nicolaus II Bernoulli through Johann Bernoulli
  - Johann II Bernoulli through Johann Bernoulli

---

#### Relation Percentage unknown

<table>
<thead>
<tr>
<th>Relation</th>
<th>All 3M</th>
<th>Top 100K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession</td>
<td>68%</td>
<td>24%</td>
</tr>
<tr>
<td>Place of Birth</td>
<td>71%</td>
<td>13%</td>
</tr>
<tr>
<td>Nationality</td>
<td>75%</td>
<td>21%</td>
</tr>
<tr>
<td>Education</td>
<td>91%</td>
<td>63%</td>
</tr>
<tr>
<td>Spouses</td>
<td>92%</td>
<td>68%</td>
</tr>
<tr>
<td>Parents</td>
<td>94%</td>
<td>77%</td>
</tr>
<tr>
<td>Children</td>
<td>94%</td>
<td>80%</td>
</tr>
<tr>
<td>Siblings</td>
<td>96%</td>
<td>83%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>99%</td>
<td>86%</td>
</tr>
</tbody>
</table>

**as of 2014**
Weakly Sup. Knowledge Extraction
(West et al. 2014)

R = parents

- Extract relation templates/queries
- parent of __
- __'s parent
- __ father
- mother of __
- ...

search logs
QA system
Weakly Sup. Knowledge Extraction
(West et al. 2014)

R = parents

Q = Frank Zappa

Explain relation templates/queries

parent of __
__’s parent
__’s father
mother of __
...

Score entities in result snippets & aggregate

Mothers of Inversion
Ray Collins
Rose Marie Colimore
Francis Zappa
Gail Zappa
Rose Marie
Does it Work?
Does it Work?
The Resource Trade-off

High quality annotations

Crowd-sourced

Auto resources
The Resource Trade-off

High quality annotations

Crowd-sourced

Auto resources
Top-level Conclusion

Syntax

Semantics
Thanks