Some Important Problems in Natural Language Processing

Mark Steedman

With Michael Auli, Jason Baldridge, Lexi Birch, Prachya Boonkwan, Johan Bos, Çem Boszahin, Ruken Çakıcı, Chris Christodoulopoulos, Stephen Clark, Greg Coppola, James Curran, Aciel Esky, Chris Geib, Julia Hockenmaier, Tom Kwiatkowski, Mike Lewis, Kira Mourão, Ron Petrick, Emily Thomforde, Mark Wilding, Luke Zettlemoyer, *and others*

Edinburgh Informatics Hamming Seminar 27 Oct 2010

Richard Hamming

- Hamming Numbers $2^x 3^y 5^z$ (Manhattan Project)
- Hamming Distance (Manhattan Distance)
- Hamming Problems

Richard Hamming

- Hamming Numbers $2^x 3^y 5^z$ (Manhattan Project)
- Hamming Distance (Manhattan Distance)
- Hamming Problems
  - What are the really important problems in your field?

Richard Hamming

- Hamming Numbers $2^x 3^y 5^z$ (Manhattan Project)
- Hamming Distance (Manhattan Distance)
- Hamming Problems
  - What are the really important problems in your field?
  - What are you working on?

Richard Hamming

- Hamming Numbers $2^x 3^y 5^z$ (Manhattan Project)
- Hamming Distance (Manhattan Distance)
- Hamming Problems
  - What are the really important problems in your field?
  - What are you working on?
  - *Why are you looking for your keys under the streetlight when you know you dropped them at the Dark End of the Street?*

A Brief History of Problems

- Until the mid-sixties, Computational Linguistics was greatly focused on finite state methods and the problem of Machine Translation. The machines were tiny, slow, and expensive.
- The Georgetown Experiment 1956-1966.
- Cyclic Translation demo: English Russian English

  Time flies like an arrow $\Rightarrow$ Time flies enjoy arrows

A Brief History of Problems

- In the aftermath of the ALPAC report (1966), there was widespread agreement that the important problems were essentially linguistic. The emphasis was on full syntactic analysis, and the need for semantically-based "understanding" to resolve the huge degree of ambiguity that the earlier work had revealed. Entire research groups were running on a single 10KHz processor and 1Mb of RAM (or less)
- By the mid '80s (by which time Moore's Law had put several times that computing power on every desktop), it was clear that the grammars and parsing techniques developed in the previous decade were not going to scale to wide-coverage.
  - The grammars were too big to manage and mainly consisted of exceptions.
  - The degree of ambiguity was too great for exhaustive search.
  - Linguistic semantics compounded the problem of ambiguity and failed to support inference and understanding.

Where we are now

- Meanwhile, some novel algorithms for statistical model estimation had been discovered. It became clear that finite-state methods scaled much better than more powerful methods, and that the most practical solution to the problem of syntactic ambiguity was to combine standard parsing algorithms with probabilistic or information-theoretic models of their yield, derived from counts of their components in human-labeled corpora or tree-banks—i.e. "supervised" machine learning.
- A major boost to this method came from the discovery that specific word-dependencies, as between a verb and the noun heading its subject, were particularly informative. Interestingly, such head-word dependency models can be seen as approximating a model of semantic predicate-argument relations.
- As an unexpected bonus, it turned out to be easier to derive large grammars automatically from treebanks than to engineer them by hand.

Where we are now

- In the present state of natural language processing research, statistical models are ubiquitous. Together with the exponential increase in computing power under Moore's Law, they have driven the remarkable progress of the last 40 years in automatic speech recognition (ASR), information retrieval (IR), and statistical machine translation (SMT).
- The most successful methods use **Supervised Learning** from data labeled by humans.
  - For parsers, the data are sets of sentences laboriously annotated with syntactic trees or dependency graphs, the largest of which are currently around 1M words in size.
  - For SMT systems, the training data are parallel text produced by human translators, of which the largest set available is around 200M words.
  - For ASR it seems to be a few thousand hours of transcribed human speech.

The State of the Art: Arabic-English SMT

The State of the Art: Arabic-English SMT

- From Al Jazeera: Arabic human translation of Reuters newswire in English:

The State of the Art: Arabic-English SMT

- From Al Jazeera: Arabic human translation of Reuters newswire in English:

ويقوم فرانز أوتش وهو ألماني يقود جهود غوغل الخاصة بالترجمة بتغذية الحاسوب بمئات الملايين من الكلمات من نصوص موازية مثل العربية والإنجليزية مستخدما وثائق الأمم المتحدة والاتحاد الأوروبي مصادر رئيس.

وعن طريقة الترجمة الجديدة قال أوتش إنه رغم أن الجودة لن تكون كاملة يعد ذلك تطورا في المساعي السابقة الخاصة بالترجمة الآلية، وإن الترجمة الصحيحة في أغلبها قد تكون جيدة بما يكفي لبعض المهام. وذكر أنه كلما زادت البيانات التي يتم تغذية النظام بها كانت النتائج أفضل.

وأثنى مايلز أوسبورن الأستاذ بجامعة أدنبره الذي قضى العام الماضي في العمل في مشروع غوغل على جهود الشركة، غير أنه لفت إلى أن البرمجيات لن تتغلب على البشر في الترجمات الماهرة كما تفعل في لعبة الشطرنج وأنه ينبغي استخدام البرمجيات للفهم وليس لإنجاز وثائق.

The State of the Art: Arabic-English SMT

The State of the Art: Arabic-English SMT

- The German Franz Ouch which leads efforts Google translation computer feeds hundreds of millions of words of parallel texts such as Arabic, English, using documents of the United Nations and the European Union key sources.

The State of the Art: Arabic-English SMT

- The German Franz Ouch which leads efforts Google translation computer feeds hundreds of millions of words of parallel texts such as Arabic, English, using documents of the United Nations and the European Union key sources.
- And how a new translation Ouch said that although the quality would not be complete That was a good in the previous translation mechanism, and that the correct translation mostly might be good enough for some tasks. He stated that more data be fed by the results were better.

The State of the Art: Arabic-English SMT

- “The German Franz Ouch which leads efforts Google translation computer feeds hundreds of millions of words of parallel texts such as Arabic, English, using documents of the United Nations and the European Union key sources.
- And how a new translation Ouch said that although the quality would not be complete That was a good in the previous translation mechanism, and that the correct translation mostly might be good enough for some tasks. He stated that more data be fed by the results were better.
- . . . He commended Miles Osborne Professor at the University of Edinburgh, who died last year at work in the company’s efforts to Google, but he pointed out that the software will not prevail over people skilled in translations as they do in the game of chess and should use software to understand and not to complete documents.”

The State of the Art: Arabic-English SMT

- “The German Franz Ouch which leads efforts Google translation computer feeds hundreds of millions of words of parallel texts such as Arabic, English, using documents of the United Nations and the European Union key sources.
- And how a new translation Ouch said that although the quality would not be complete That was a good in the previous translation mechanism, and that the correct translation mostly might be good enough for some tasks. He stated that more data be fed by the results were better.
- . . . He commended Miles Osborne Professor at the University of Edinburgh, who died last year at work in the company’s efforts to Google, but he pointed out that the software will not prevail over people skilled in translations as they do in the game of chess and should use software to understand and not to complete documents.”

The arabic words for “passed” and “died” are homographs, and the arabic news-data model favors the latter

2007: English-Arabic-English

2007: English-Arabic-English

- Time flies like an arrow.

  الوقت الذباب يشبه السهم.

  time-DEF flies-DEF resemble arrow-DEF

  "Time flies like arrow."

2007: English-Arabic-English

- Time flies like an arrow.
  الوقت الذباب يشبه السهم.
  time-DEF flies-DEF resemble arrow-DEF
  "Time flies like arrow."

  Fruit flies like a banana.
  ذباب الفاكهة مثل الموز.
  Flies-N fruit-N resemble banana-N
  "Fruit flies like bananas."

Soon after I published these results in 2007, some of these specific examples were fixed by Google. However, there are still plenty more like them currently out there for you to find for yourself.

The View from the Long Tail

The View from the Long Tail

- This is the bank that bought the company.
  وهذا البنك هو ان اشترت الشركة.
  "This is the bank that bought the company."

The View from the Long Tail

- This is the bank that bought the company.

  وهذا البنك هو ان اشترت الشركة.

  "This is the bank that bought the company."

  This is the company that the bank bought.

  هذه هي الشركة التي اشترت البنك.

  "*This is the company that bought the bank."

The View from the Long Tail

- This is the bank that bought the company.

  وهذا البنك هو ان اشترت الشركة.

  "This is the bank that bought the company."

  This is the company that the bank bought.

  هذه هي الشركة التي اشترت البنك.

  "*This is the company that bought the bank."

  This is the bank that wants to buy the company.

  هذا هو المصرف الذي يريد لشراء الشركة.

  "This is the bank, which wants to buy the company."

The View from the Long Tail

- This is the bank that bought the company.
  وهذا البنك هو ان اشترت الشركة.
  "This is the bank that bought the company."

  This is the company that the bank bought.
  هذه هي الشركة التي اشترت البنك.
  "*This is the company that bought the bank."

  This is the bank that wants to buy the company.
  هذا هو المصرف الذي يريد لشراء الشركة.
  "This is the bank, which wants to buy the company."

  This is the company which the bank wants to buy.
  هذه هي الشركة التي تريد شراء البنك.
  "*This is the company that wants to buy the bank."

More Long Tail

More Long Tail

- This is the company that said the bank bought bonds.

  هذه هي الشركة التي قال البنك بشراء السندات.

  "This is the company that said the bank bought the bonds."

More Long Tail

- This is the company that said the bank bought bonds.

  هذه هي الشركة التي قال البنك بشراء السندات.

  "This is the company that said the bank bought the bonds."

  This is the company that the bank said bought bonds.

  هذه هي الشركة التي قال البنك بشراء السندات.

  "*This is the company that said the bank bought the bonds."

More Long Tail

- This is the company that said the bank bought bonds.
هذه هي الشركة التي قال البنك بشراء السندات.
"This is the company that said the bank bought the bonds."

  This is the company that the bank said bought bonds.
هذه هي الشركة التي قال البنك بشراء السندات.
"*This is the company that said the bank bought the bonds."

  These are the bonds that the company said that the bank bought.
هذه هي سندات الشركة أن البنك اشترى.
"*These are the bonds that the bank bought the company."

Who cares?

- *Not surprisingly*, SMT is bad at non-subject relatives. So what?
- There are around 1000 object relatives in the Penn Treebank.
- Getting them right isn't going to significantly affect your global dependency-recovery rate or Bleu score.
- However, they are semantically crucial
- Genres like Questions have higher rates.
- Moreover, the more of the easy stuff we get right, the more this bad stuff will matter.
  - Q: What do frogs eat?
  - A: Herons.
- This is somewhat like what our colleagues in animation call the "uncanny valley".

What to Do?

- Keep on looking under the streetlight:
  - Give thanks for Moore's Law.
  - Get more data, and hit it with the latest fashion in statistical models.
- This may not work:
  - Accuracy in most areas (WER in ASR, Bleu score in SMT, Eval-b for parsers) is *at best* linear in the logarithm of the training data.
  - Extrapolation of learning curves suggests impractical data requirements (Knight and Koehn 2004; Moore 2003; Lamel, Gauvain and Adda 2002).
    - No-one is going to give us even 10M words of Penn treebank
    - We can't wait around for 2BN words of parallel text
    - The amount of speech data that would be required to bring HMM ASR up to human standard seems to be about 1M hours.
- Unsupervised learning of such systems from unlabeled data doesn't seem to work.

Interpolating Higher Level Information

- There is every indication that high level information from syntax and semantics will help with this problem (Hassan, Sima'an and Way 2009; Birch, Osborne and Koehn 2007).
- This claim is very hard to prove, because for applications like ASR and SMT the syntax and semantics needs to be **incremental** for this to work (Roark 2001).
- Most of the available theories, including treebank grammars, lack this property.
- Nevertheless, even speakers of verb-final languages like Japanese are convinced that their interpretation is incremental.

The Hamming Alternative

- "What are the most important problems in your field?"
  - Breaking the asymptote of approximate approaches to syntax and semantics.
  - Building the right kind of grammar at a large enough scale for reliable parsers to support QA, IR, SMT etc.
  - Building statistical models large enough to drive those parsers.
  - Building a semantics that will support practical inference beyond the sentence for those purposes.
  - Doing all of this using unlabeled data.

Problem 1: Fix the Grammars

- Most grammars (Chomsky *passim*, GPSG, HPSG, LFG) are **rule-based**:

  $NP \rightarrow Det \quad Noun$

- but there are always **exceptions**:

  $[[\text{Whisky}]_{Noun}\ [\text{galore}]_{Det}]_{NP}$

So **lexicalize** the grammar (TAG, CCG):

$\text{a}, \text{the}, \text{every} := NP/N \qquad \text{galore} := NP\backslash N$

Restrict rules to **adjacent** operators with as much incrementality as possible (Steedman 2000)

Problem 2: Learn from Unlabeled Data

- Pure unsupervised learning is too hard.
- Partially unsupervised learning by generalizing from supervised grammars may be possible.
  - Sentences in which we know everything with high confidence except one word such as “galore” might allow us to bootstrap a lexical entry for the unseen word which allows an analysis with high probability (Thomforde 2008).
  - Sentences in which we think we know everything but the model says every parse is low probability might allow us to bootstrap a new lexical entry such as causative transitive “walk” for seen intransitive “walk”.
  - Such methods might be boosted with tiny amounts of language-specific data (Boonkwan).

Estimating the model for new lexical items is the hard part.

Problem 3: Get Labeled Data Automatically

- Children learn language with great facility from paired strings and (noisy, confounded) contextually available meanings, learning a parsing model for universal grammar in much the same way as a supervised parser (Kwiatkowksi, Goldwater and Steedman 2009).
- Human operators like travel agents map queries onto database queries and database returns onto answers. Can we learn from these data (e.g. ATIS Zettlemoyer and Collins 2007)?

  show me information on american airlines from fort worth texas to philadelphia

  $\lambda x.airline(x, americanairlines) \wedge from(x, fortworth) \wedge to(x, philadelphia)$

- Kwiatkowksi et al. (2010) shows how to do this using Higher Order Unification to produce all possible decompositions of the logical form paired with all possible substrings of the sentence.

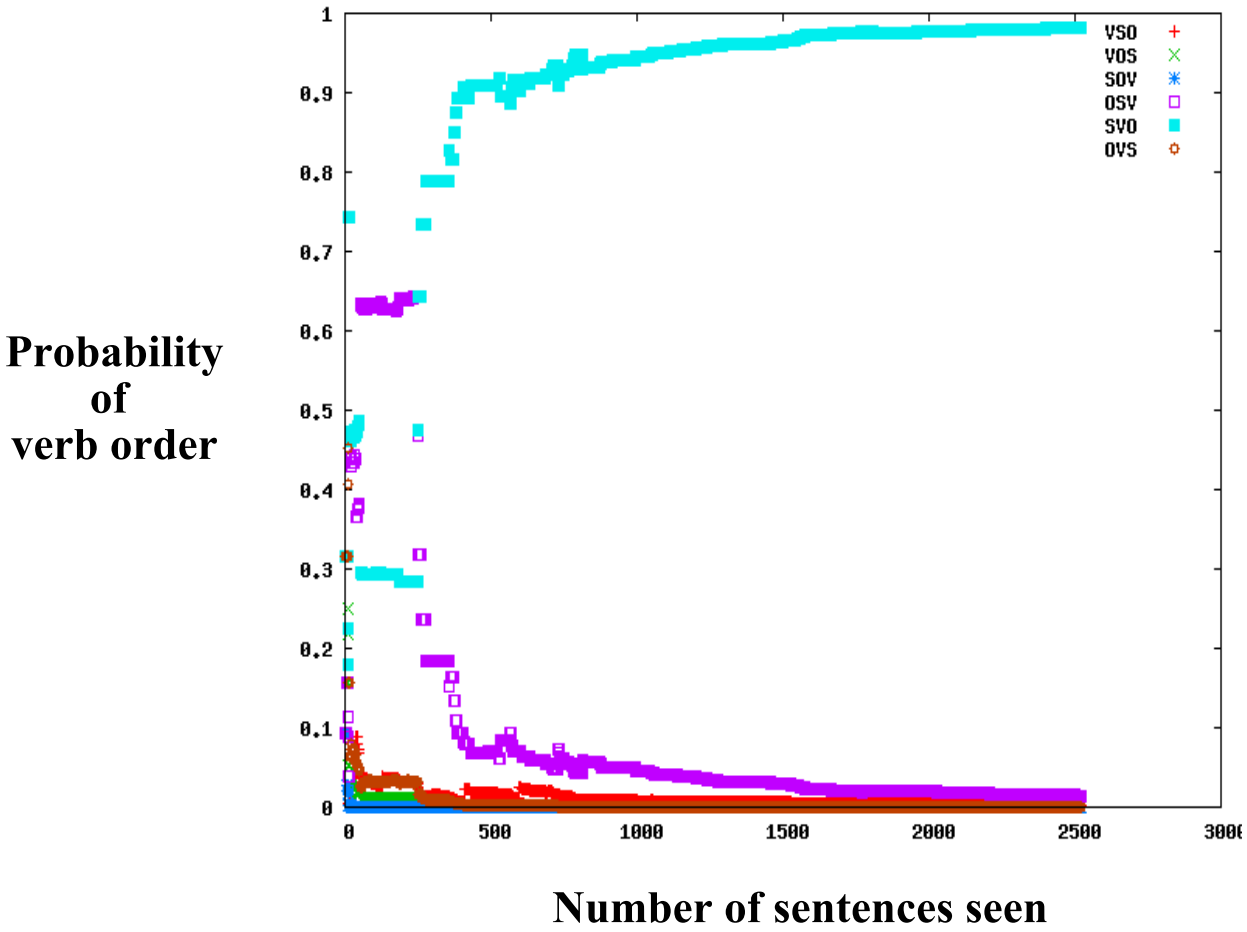

Figure 1: Learning SVO word order from the CHILDES dependency bank of child-directed utterance using Variational Bayes (Kwiatkowksi, Goldwater and Steedman 2009).

Problem 4: Fix the Semantics

- Understanding language always involves inference beyond the literal meaning of what is said.
- Most semantics is expressed in some version of first order logic.
- Such interpretations are often non-isomorphic to syntax, as below, where the object has scope over the subject:

  A silencer must be fitted to every vehicle

- Multiple equivalent interpretations abound: Koller and Thater (2006) note that in one popular version the following has 3960 distinct interpretations all of which are equivalent:

  For travelers going to Finnmark there is a bus service from Oslo to Alta through Sweden.

Problem 5: Fit the Semantics to Shallow Inference

- Logical form should be lexicalized, and projected monotonically by the same adjacent operations as syntactic category (Lewis).
- As many expressions like “a silencer” as possible should be replaced by in situ dependent or independent individual descriptions to maintain isomorphism between logical form.
- Shallow inference on the basis of taxonomies like WordNet is dependent on polarity, so polarity should be directly represented in the logical language (cf. MacCartney 2009):

  Emmylou doesn’t keep a dog

  $\models$ Emmylou doesn’t keep a poodle

  $\not\models$ Emmylou doesn’t keep an animal

Efficient Representation

- A representative of every company saw a sample

$$\forall y \left[company' y \rightarrow saw' \left(\left\{ \begin{array}{c} sk^{(y)} \\ sk \end{array} \right\} sample' \right) \left(\left\{ \begin{array}{c} sk^{(y)} \\ sk \end{array} \right\} \lambda x.representative' x \wedge of' yx \right) \right]$$

a. $\forall y[(company' y \wedge of' y\ sk^{(y)}_{representative'}) \rightarrow saw' sk^{(y)}_{sample'} y]$

b. $\forall y[(company' y \wedge of' y\ sk_{representative'}) \rightarrow saw' sk^{(y)}_{sample'} y]$

c. $\forall y[(company' y \wedge of' y\ sk^{(y)}_{representative'}) \rightarrow saw' sk_{sample'} y]$

d. $\forall y[(company' y \wedge of' y\ sk_{representative'}) \rightarrow saw' sk_{sample'} y]$

Problem 6: Identify the Universal Language of Thought

- Since a universal semantics must be directly hung onto a universal embodied animal cognition (to which we have no access), and children can hang any language directly onto that semantics, we should keep the elements of the logical language as close to the elements of Universal Grammar as we can.
- Linguists aren't being as helpful as they might be in telling us what UG is.
- Could we machine-learn the elements from parallel text in lots of more analytic languages?
- Say by unsupervised clustering of parts of speech (Christodoulopoulos, Goldwater and Steedman 2010) and mapping CCG categories from English.
- Soon we should be into a virtuous cycle, where we can use our parser to build more powerful resources than WordNet automatically, and let it loose to read the web for us, because all the important problems have been solved.

Thanks!

Mark Steedman

With Michael Auli, Jason Baldridge, Lexi Birch, Prachya Boonkwan, Johan Bos, Çem Boszahin, Ruken Çakıcı, Chris Christodoulopoulos, Stephen Clark, Greg Coppola, James Curran, Aciel Esky, Chris Geib, Julia Hockenmaier, Tom Kwiatkowski, Mike Lewis, Kira Mourão, Ron Petrick, Emily Thomforde, Mark Wilding, Luke Zettlemoyer, *and others*

Birch, Alexandra, Miles Osborne, and Philipp Koehn. 2007. "CCG SuperTags in Factored Translation Models." In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, 9–16. ACL.

Christodoulopoulos, Christos, Sharon Goldwater, and Mark Steedman. 2010. "Two Decades of Unsupervised POS tagging—How Far Have We Come?" In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 575–584. ACL.

Hassan, Hany, Khalil Sima'an, and Andy Way. 2009. "A Syntactified Direct Translation Model with Linear-time Decoding." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1182–1191. Singapore: Association for Computational Linguistics.

Knight, Kevin, and Philipp Koehn. 2004. "What's New in Statistical Machine Translation." Tutorial, HLT/NAACL.

Koller, Alexander, and Stefan Thater. 2006. "An Improved Redundancy

Elimination Algorithm for Underspecified Descriptions." In *Proceedings of COLING/ACL-2006*. Sydney.

Kwiatkowksi, Tom, Sharon Goldwater, and Mark Steedman. 2009. "Computational Grammar Acquisition from CHILDES data using a Probabilistic Parsing Model." In *Workshop on Psycho-Computational Models of Human Language Acquisition, at the 31st Annual Meeting of the Cognitive Science Society*.

Kwiatkowksi, Tom, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. "Inducing Probabilistic CCG Grammars from Logical Form with Higher-Order Unification." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1223–1233.

Lamel, Lori, J-L. Gauvain, and G. Adda. 2002. "Unsupervised Acoustic Model Training." In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 877–880. IEEE.

MacCartney, Bill. 2009. *Natural Language Inference*. Ph.D. thesis, Stanford University.

Moore, Roger. 2003. "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners." In *Proceedings of Eurospeech Conference*, 2582–2585.

Roark, Brian. 2001. "Probabilistic top-down parsing and language modeling." *Computational Linguistics*, 27, 249–276.

Steedman, Mark. 2000. *The Syntactic Process*. Cambridge, MA: MIT Press.

Thomforde, Emily. 2008. *Semi-Supervised Lexical Acquisition for Wide-Coverage Parsing*. Ph.D. thesis, University of Edinburgh. In preparation.

Zettlemoyer, Luke, and Michael Collins. 2007. "Online Learning of Relaxed CCG Grammars for Parsing to Logical Form." In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*, 678–687. ACL.