Why NLP?

Natural Language Processing
CS 6120—Spring 2014
Northeastern University

David Smith
Codes
Completely Automated Public Turing test to tell Computers and Humans Apart
Fine, walk away. I'm gonna go cry into a pint of Ben&Jerry's Brownie Batter(tm) ice cream [link], then take out my frustration on a variety of great flash games from PopCap Games(r) [link].
TENSE AND MOOD IN INDO-EUROPEAN SYNTAX*

1. THE HISTORICAL PRESENT

The 'historical' or 'dramatic' present tense used in narrating past events, which is common in many Indo-European languages, has always been interpreted in essentially semantic terms. A typical traditional formulation is it is quite mistaken to transfer it to the earlier stages of Indo-European. In Greek, Old Irish, and Old Norse, for example, the historical present has quite different syntactic and semantic properties, to which the traditional idea, or any of its variants\(^2\), must utterly fail to do justice.

* This work was supported in part by the Joint Services Electronics Program under Contract DA36-039-AMC-03200(E); in part by the National Science Foundation (Grant GP-2495), the National Institutes of Health (Grant MH-04737-05), the National Aeronautics and Space Administration (Grant NsG-496), and the U.S. Air Force (ESD Contract AF 19 (628)-2487). – I thank Michael Connolly, Eric Hamp, Einar Haugen, George Lakoff, Calvert Watkins, and Roy Wright for offering valuable criticism and/or referring me to some of the examples cited here.
One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

Warren Weaver to Norbert Wiener
4 March 1947
ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO
THE ENTSCHEIDUNGSPROBLEM

By A. M. Turing.

[Received 28 May, 1936.—Read 12 November, 1936.]

The "computable" numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable numbers, it is almost equally easy to define and investigate computable functions of an integral variable or a real or computable variable, computable predicates, and so forth. The fundamental problems involved are, however, the same in each case, and I have chosen the computable numbers for explicit treatment as involving the least cumbrous technique. I hope shortly to give an account of the relations of the computable numbers, functions, and so forth to one another. This will include a development of the theory of functions of a real variable expressed in terms of computable numbers. According to my definition, a number is computable if its decimal can be written down by a machine.

In §§ 9, 10 I give some arguments with the intention of showing that the

with the m-configuration written below the scanned symbol. The successive complete configurations are separated by colons.

\[\ldots\]

This table could also be written in the form

\[\ldots\]
The Turing Test

Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.
Modularity
Linguistic Modules

• Phonetics and phonology
• Morphology
• Syntax
• Semantics
• Pragmatics
• Discourse

• With lots of crossings between levels!
Phonetics and Phonology

- Phonetics: language sounds & their physiology
- Phonology: systems of discrete sounds in languages
  - E.g.: devoicing of *it is* to *it’s*
  - E.g.: syllable structure: *sign, signify*
Morphology

• Inflectional (in some languages):
  • love $\rightarrow$ loved

• Derivational:
  • tea-cup, un-helpful, with-stand, craisin

• Turkish: uygarlastirimadiklarimizdanmissinizcasina
  • uygar las tir ama dik lar imiz dan mis siniz casina

• (behaving) as if you are among those whom we could not civilize
Morphological Tagging

There are many kinds of trench mortars.

c. Klimatizovaná jídelna, světlá místnost pro snídaně.

Air-conditioned dining room,
Syntax

IP

NP

VP

V

NP

language

Constituency

Speaking

changes

language

Dependency

Verbing

weirds

language

SUBJ

OBJ
Semantics

Pierre Vinken, 61 years old, will join the board as a nonexecutive director

PropBank join predicate

<table>
<thead>
<tr>
<th>ARG0</th>
<th>ARG1</th>
<th>ARG-PRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vinken</td>
<td>board</td>
<td>director</td>
</tr>
</tbody>
</table>
Pragmatics

• Context affects meaning
• Conversational implicature
  • *May I speak to your mother? Yes.*
• Speech acts: “how to do things with words”
  • *I grant you permission to speak.*
Discourse

- Study of units larger than a single utterance
- Turn taking
- Coreference
- Organized exposition
It All Hangs Together

Inflectional Morphology

Phonology

Syntax

Pragmatics

Discourse

Semantics

Derivational Morphology

I've loved our online chats these past few months, Lisa. Me too. I really like you, Rob.

It's just... now and then you mention products you like, and... I worry. What? Honey...

Before this goes any further, I think we should go get tested. You know, together. You don't trust me? I just want to be sure.

Okay, mine says "library". Yours? Oh God. I'm more than a spambot! Our love was real.

Goodbye, Lisa.
Applications
<bookstore>
  <book category="COOKING">
    <title lang="en">Everyday Italian</title>
    <author>Giada Di Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>

  <book category="COOKING">
    <title lang="en">Empire of the Son</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>

  <book category="WOMEN"
    <title lang="en">The Inner Circle</title>
    <author>Erik T. Erickson</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
Geometry, Euclidis. 16 Et hic quidem punctus, centrum circuli dicitur. 17 Diameter circuli, est linea recta, qua super eius centrum transiens, extremitateque suas curvaturae applicat, circuli in duo media diuisit. 18 Semi-circulus, est figura plane diametro circuli, et mediatur curvaturae conica. 19 Portio circuli, est figura plana, recta linea, & parte curvaturae consens, semi-circulo quam a maior aut minor.

Semicircle. Minor partio. Rectilinear figura sunt, quae rectis lineis continantur. 21 Quorum quaedam tria latera, quae tribus rectis lineis sunt. 22 Quaedam quadrilatera, quatuor rectis lineis sunt. 23 Quaedam quadrilatera, quae pluribus quatuor rectis lineis continentur. 24 Figurarum tria latera, alia est trianguulus, habet tria latera aequalia. 25 Alius triangulus, duo habet aequalia latera. 26 Alius triangulus trii inaequalia lateri. 27 Habent iuribus alia est orthogonium, unum rectum cum aequalia latera habet. 28 Est orthogonium, quae quadrilaterum anguli habent. 29 Alius est orthogonium, quae tres anguli sunt acuti. 30 Est orthogonium, quae quadrilaterum anguli lateri. 31 Alius est orthogonium, quae quadrilaterum anguli lateri. 32 Alius est orthogonium, quae quadrilaterum anguli lateri. 33 Alius est orthogonium, quae quadrilaterum anguli lateri.

Table 2: Unlabeled dependency accuracy for German and Spanish with different training conditions and training set sizes.

<table>
<thead>
<tr>
<th>Baselines</th>
<th>Dependency accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modify prev.</td>
<td>18.2</td>
</tr>
<tr>
<td>Modify next</td>
<td>27.5</td>
</tr>
<tr>
<td>Training sentences</td>
<td>1k 10k 1k 10k</td>
</tr>
<tr>
<td>EM</td>
<td>30.2 30.8 25.6 24.9</td>
</tr>
<tr>
<td>Hard proj.</td>
<td>66.2 64.8 59.1 60.1</td>
</tr>
<tr>
<td>Hard proj. w/EM</td>
<td>58.6 59.8 53.0 52.8</td>
</tr>
<tr>
<td>QG w/EM</td>
<td>68.5 66.9 64.8 64.8</td>
</tr>
</tbody>
</table>

To find sentences where enough links were projected to completely determine a target language tree. Of course, we needed to filter more than 1000 sentences of bi.tree to output 1000 training sentences in this way. With this subset, we can simply perform supervised training. As discussed in §2, these links are still quite noisy. Performance in fact suffers when we add more of this noisy training data. Still, this method is a substantial improvement over the baseline and unsupervised EM.

Instead of finding fully projected trees, we can simply take the one-to-one projected links are given, impute expected counts for the remaining structures with EM, and update our models. This approach (“hard projection with EM”), however, performed worse than using only the fully projected trees. In fact, only the first iteration of EM with this method made any improvement. Afterwards, EM degraded accuracy further from the numbers in table 2.

5.2 Unsupervised Learning

5.4 QG Projection
Er wird in den Strassen wandern

He will in the streets walk
He will walk in the streets

Er wird in den kleinen Strassen wandern

He will in the small streets walk
He is in the small streets hike
Who is the leader of France?

Henri Hadjenberb, who is the leader of France’s Jewish community.

Bush met with French president Jacques Chirac.
Multilingual “Topics”
European Parliament Corpus

| DA | børn familie udnyttelse børns børnene seksuel |
| DE | kinder kindern familie ausbeutung familien eltern |
| EL | παιδιά παιδίων οικογένεια οικογένειας γονείς παιδικής |
| EN | children family child sexual families exploitation |
| ES | niños familia hijos sexual infantil menores |
| FI | lasten lapsia lapset perheen lapsen lapsiin |
| FR | enfants famille enfant parents exploitation familles |
| IT | bambini famiglia figli minori sessuale sfruttamento |
| NL | kinderen kind gezin seksuele ouders familie |
| PT | crianças família filhos sexual criança infantil |
| SV | barn barnen familjen sexuellt familj utnyttjande |

| DA | mål nå målsætninger målet målsætning opnå |
| DE | ziel ziele erreichen zielen erreicht zielsetzungen |
| EL | στόχους στόχο στόχος στόχων στόχοι επίτευξη |
| EN | objective objectives achieve aim ambitious set |
| ES | objetivo objetivos alcanzar conseguir lograr estos |
| FI | tavoite tavoitteet tavoitteena tavoitteiden tavoitteenan |
| FR | objectif objectifs atteindre but cet ambitieux |
| IT | obiettivo obiettivi raggiungere degli scopo quello |
| NL | doelstellingen doel doelstelling bereiken bereikt doelen |
| PT | objectivo objectivos alcançar atingir ambicioso conseguir |
| SV | mål målet uppnå målen målsättningar målsättning |

| DA | andre anden side ene andet øvrige |
| DE | anderen andere einen wie andererseits anderer |
| EL | άλλες άλλα άλλη άλλων άλλους όπως |
| EN | other one hand others another there |
| ES | otros otras otro otra parte demás |
| FI | muiden toisaalta muita muut muihin muun |
| FR | autres autre part côté ailleurs même |
| IT | altri altre altro altra dall parte |
| NL | andere anderzijds anderen ander als kant |
| PT | outros outras outro lado outra neutros |
| SV | andra sidan å annat ena annan |
Multilingual “Topics” Wikipedia comparable articles
world ski km won
actor role television actress
ottoman empire khan byzantine
Projecting Hidden Structure

Annotations From Existing English Tools

[PLACE]

[National laws] applying in [Hong Kong]

[PLACE]

Induced Annotations for Chinese
NLP Tasks

• Analog to digital
  • OCR, Speech Recognition
• Individual language modules
  • Morphology, Syntax, Semantics, and Discourse
• Language to data
  • Information extraction and retrieval
• Language to language
  • Translation, summarization, dialogue systems
Monolingual & Multilingual

- Analysis technologies for languages
  - Morphology, syntax, semantics
- Translation technologies
  - Dictionaries, cross-lingual IR, MT
- Multilingual exploratory data analysis
  - Clustering, classification → model building
A Few Problems
Morphological Ambiguity

There are many kinds of trench mortars.

c. Klimatizovaná jídelna, světlá místnost pro snídaně.

Air-conditioned dining room,
Syntactic Ambiguity

**S**

**NP**

**NNP**

Fed

**VP**

raises

**NP**

interest

rates

**S**

**NP**

**N**

Fed

raises

**NP**

interest

rates

**S**

**NP**

Fed

raises

**NP**

interest

rates

**NP**

Fed

raises

**NP**

interest

rates

CD

N

0.5

%
More Ambiguity

- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Ban on Nude Dancing on Governor’s Desk
Why is NLP Hard?

- The rules are ambiguous
- We don’t know the rules
- We need to combine lots of weak evidence
- It’s AI complete
- Language is nearly co-extensive with humanity
- To the rescue: probability, machine learning
Why is NLP in CS?

- How about...
- Linguistics
- Statistics
- Psychology and Cognitive Science
- The Lang/Lit Humanities
- All of the above!
- Focus on algorithms, data analysis, engineering
What You’ll Learn in NLP

• Looking at data
• Phenomena and problems
• Modeling data
• Linguistic and statistical tools
• Algorithms and implementation
• Efficient computation, practical systems
No Really, What’ll I Learn?

- Models of language
  - n-grams, grammars, generative, discriminative
- Algorithms to tame complexity
  - Finite-state models and regular expressions
- Context-free grammars and parsers
- Problem solving: classification, structured prediction, translation
Who – Where – When

• Instructor: David Smith
  • dasmith@ccs.neu.edu
  • WVH 356, Th 3-5 or by appointment
• TA: Moonyoung Kang
  • yerihyo@ccs.neu.edu
  • WVH 472, Tu 3-5
• Thursdays 6-9, Hayden 221
  • www.ccs.neu.edu/course/6120sp14
What

- Graduate course in NLP
  - Learning to *read papers* in NLP
- Discussion and participation (20%)
- Homework assignments (4 for 40%)
- Literature review (40%)
What

• Lectures introduce algorithms, models, learning methods

• Background reading in two books:
  • *Speech and Language Processing*. Jurafsky & Martin
  • *Linguistic Structure Prediction*. Noah Smith (no relation, but sometime coauthor)
Thanks