NLP: Going for low-hanging fruit

Bonnie Webber (University of Edinburgh)

June 19, 2012

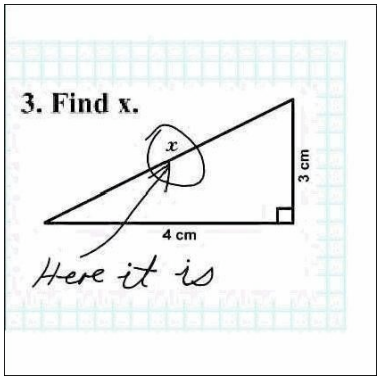
1. Introduction
   - NL offers many hard problems
   - But NL features mean low-hanging fruit as well
2. Some low-hanging fruit in computational discourse
   - Text segmentation
   - Recognizing coherence relations
3. Conclusion

Hard problems in NL: Pragmatic inference

Pragmatic inference aims to account for **what** the speaker is saying or asking.

Hard problems in NLP: Pragmatic inference

What is the speaker asking?



Pragmatic inference is a hard problem.

Hard problems in NLP: Intention recognition

Intention recognition aims to identify **why** the speaker is telling or asking something of the listener.

Hard problems in NLP: Intention recognition

Intention recognition aims to identify why the speaker is telling or asking something of the listener.

Why are you telling me?

Hard problems in NLP: Intention recognition

Intention recognition aims to identify why the speaker is telling or asking something of the listener.

Why are you telling me?

"My New Philosophy" From *You're a Good Man, Charlie Brown*

Intention recognition is a hard problem.

Hard problems in NLP: Recognizing coherence relations

Coherence relation recognition aims to identify the **connection** between two sentences.

(1) Don't worry about the world coming to an end today.

Hard problems in NLP: Recognizing coherence relations

Coherence relation recognition aims to identify the **connection** between two sentences.

(2) Don't worry about the world coming to an end today.
It is already tomorrow in Australia.
[Charles Schulz]

Hard problems in NLP: Recognizing coherence relations

Coherence relation recognition aims to identify the **connection** between two sentences.

(3) Don't worry about the world coming to an end today. [**reason**]
It is already tomorrow in Australia.
[Charles Schulz]

(4) I don't make jokes.
I just watch the government and report the facts.
[Will Rogers]

Hard problems in NLP: Recognizing coherence relations

Coherence relation recognition aims to identify the **connection** between two sentences or clauses.

(5) Don't worry about the world coming to an end today. [**reason**]
It is already tomorrow in Australia.
[Charles Schulz]

(6) I don't make jokes. [**alternative**]
I just watch the government and report the facts.
[Will Rogers]

When not explicitly marked, recognizing coherence relations is a hard problem.

Hard problems in NLP: Script-based inference

Script-based inference aims to identify aspects of events that the speaker hasn't made explicit.

(7) Four elderly Texans were sitting together in a Ft. Worth cafe. When the conversation moved on their spouses, one man turned and asked, "Roy, aren't you and your bride celebrating your 50th wedding anniversary soon?"

"Yup, we sure are," Roy replied.

"Well, are you gonna do anything special to celebrate?"

The old gentleman pondered for a moment, then replied, **"For our 25th anniversary, I took the misses to San Antonio."**

Hard problems in NLP: Script-based inference

Script-based inference aims to identify aspects of events that the speaker hasn't made explicit.

(8) Four elderly Texans were sitting together in a Ft. Worth cafe.

When the conversation moved on their spouses, one man turned and asked, "Roy, aren't you and your bride celebrating your 50th wedding anniversary soon?"

"Yup, we sure are," Roy replied.

"Well, are you gonna do anything special to celebrate?"

The old gentleman pondered for a moment, then replied, "For our 25th anniversary, I took the misses to San Antonio. **For our 50th, I'm thinking 'bout going down there again to pick her up.**"

Script-based inference is a hard problem.

Understanding Natural Language isn't easy: Negation

My own hard problem in NL is any sentence with >1 negation or quantifier.

(9) To: Mr. Clayton Yeutter, Secretary of Agriculture, Washington, D.C.

Dear sir: My friends over in Wichita Falls TX, received a check the other day for $1,000 from the government for **not** raising hogs. So, I want to go into the "**not** raising hogs" business myself.

What I want to know is what is the best type of farm **not** to raise hogs on, and what is the best breed of hogs **not** to raise? I would prefer **not** to raise Razor Back hogs, but if that is **not** a good breed **not** to raise, then I can just as easily **not** raise Yorkshires or Durocs.

Now another thing: These hogs I will **not** raise will **not** eat 100,000 bushels of corn. I understand that you also pay farmers for **not** raising corn and wheat. Will I qualify for payments for **not** raising wheat and corn **not** to feed the 4,000 hogs I am **not** going to raise?

Understanding Natural Language isn't easy

But if every problem in NL were hard, computational linguists and researchers in Language Technology would have quit long ago.

They haven't because NL also offers **low-hanging fruit**, that's easier to pick.

Where does low-hanging fruit come from?

Sources of low-hanging fruit in NLP

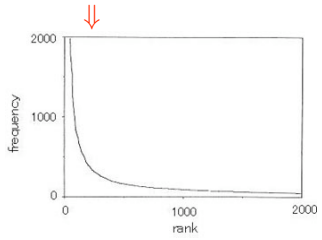
At least three (maybe four) sources of low-hanging fruit in NLP:

- Phenomena with **Zipfian distributions**;
- Availability of **low-cost proxies**;
- Acceptability of **a less than perfect** solutions;
- High value of **recall**.

N.B. Low-hanging **doesn't** mean computationally trivial: Complex algorithmic and/or statistical calculations are often involved.

Sources of low-hanging fruit (I)

In a **Zipfian distribution**, frequency varies **inversely** with rank.



This was first noticed with respect to **word tokens** in text.

The 1M-word Brown Corpus contains tokens of 39440 words.

- The top 135 words account for half the tokens ($\sim$ 500k).
- A large proportion of the 39,300 words in the **long tail** occur only once.

Sources of low-hanging fruit (I)

○ Also Zipfian is the distribution of **discourse connectives** (conjunctions, discourse adverbials) in the Penn Discourse TreeBank [Prasad *et al*, 2008], annotation over the 1M-word Penn WSJ Corpus.

Explicit Conn	No. of tokens	Explicit Conn	No. of tokens
but	3308	therefore	26
and	3000	otherwise	24
if	1223	as soon as	20
because	858	accordingly	5
while	781	if and when	3
however	465	conversely	2
...	...	...	...

Sources of low-hanging fruit (I)

○ Probably Zipfian is the distribution of **syntactic constructions** in text, although the ranking of different constructions may be genre-specific.

Zipfian distributions are a source of low-hanging fruit whenever

- the mass at the front can be handled (relatively) easily;
- the long tail can be ignored without dire consequences.

Sources of low-hanging fruit (I)

N.B. Zipfian distributions can only hold of phenomena whose tokens can be classified into **discrete categories**, whose frequency can then be counted.

That's not always possible — e.g., **animacy** — suggesting that animacy-based decisions may not be low-hanging fruit.

Sources of low-hanging fruit (II)

NL often offers **proxies** that are simpler than the full blown phenomenon:

- **Word stems**, as **proxies** for words.
- **Bag of words**, as a **proxy** for a sentence or a text.
- **Bag of sentences**, as a **proxy** for a text.
- (Probabilistic) **CFG**, as a **proxy** for a NL grammar.
- Relative **web/corpus frequency**, as a **proxy** for (relative) correctness.

Being able to exploit a good proxy, rather than the phenomenon itself, makes for low-hanging fruit.

Sources of low-hanging fruit (III)

Other sources of low-hanging fruit are **task-specific** — e.g., there's low-hanging fruit when a less-than-perfect solution is acceptable.

- Automated PoS-taggers have been used for years, even though
- The set of PoS-tags used in tagging is less-than-perfect.

In the commonly used Penn Tag Set (45 tags), **titles** (*Mr.*, *Ms.*, *Dr.*) are lumped together with singular proper nouns (**NNP**):

the Texas Rangers	the/DT Texas/NNP Rangers/NNPS
Prof. David Beaver	Prof./NNP David/NNP Beaver/NNP

even though **titles** clearly have a different distribution.

When it doesn't matter, a task can be low-hanging fruit.

Sources of low-hanging fruit (IV)

Selection tasks can be low-hanging fruit if **recall** is valued at least as much as **precision**.

Recall: The proportion of relevant items that are selected (TP/TP+FN)

Precision: The proportion of selected items that are relevant (TP/TP+FP)

Such tasks leave the real decision to the user who sees the output.

Modern search engines exploit this, in some cases **ranking** items by their likelihood of relevance.

Some low-hanging fruit in Computational Discourse

I want to turn now to some low-hanging fruit in my own area of **Computational Discourse**.

- Text segmentation
- Coherence relation recognition

in order to show that:

Even **discourse** has low-hanging fruit.

Text structure and segmentation

Texts often have an underlying **high-level structure**:

- encyclopedia articles
- news reports
- scientific papers
- transcripts of speech events (meetings, lectures, etc.)
- ...

This is what **text segmentation** aims to make explicit.

High-level structure of encyclopedia articles

	Wisconsin	Louisiana	Vermont
1	Etymology	Etymology	Geography
2	History	Geography	History
3	Geography	History	Demographics
4	Demographics	Demographics	Economy
5	Law and government	Economy	Transportation
6	Economy	Law and government	Media
7	Municipalities	Education	Utilities
8	Education	Sports	Law and government
9	Culture	Culture	Public Health
10	...	...	...

Wikipedia articles about US states

High-level structure of news reports

News reports have an **inverted pyramid** structure:

- Headline
- Lede paragraph, conveying **who** is involved, **what** happened, **when** it happened, **where** it happened, **why** it happened, and (optionally) **how** it happened
- Body, providing more detail about who, what, when, ...
- Tail, containing less important information

High-level structure of scientific papers

Scientific papers (and, more recently, their abstracts) have a high-level structure, comprising:

- **Objective** (aka *Introduction*, *Background*, *Aim*, *Hypothesis*)
- **Methods** (aka *Method*, *Study Design*, *Methodology*, etc.)
- **Results** or *Outcomes*
- **Discussion**
- Optionally, **Conclusions**

High-level structure of meetings

3 A: Good morning everybody.
4 A: Um I'm glad you could all come.
5 A: I'm really excited to start this team.
6 A: Um I'm just gonna have a little PowerPoint presentation for us, for our kick-off meeting.

7 A: My name is Rose [Anonymized].
8 A: I I'll be the Project Manager.

9 A: Um our agenda today is we are gonna do a little opening
10 A: and then I'm gonna talk a little bit about the project,
11 A: then we'll move into acquaintance such as getting to know each other a little bit, including a tool training exercise.
12 A: And then we'll move into the project plan,
13 A: do a little discussion
14 A: and close,
15 A: since we only have twenty five minutes.

16 A: First of all our project aim.
17 A: Um we are creating a new remote control which we have three goals about,
18 A: it needs to be original, trendy and user-friendly.
19 A: I'm hoping that we can all work together to achieve all three of those.
20 A: Um so we're gonna divide us up into three compa three parts.
21 A: First the functional design
22 A: which will be uh first we'll do individual work,
23 A: come into a meeting,
24 A: the conceptional design, individual work and a meeting,
25 A: and then the detailed design, individual work and a meeting.
26 A: So that we'll each be doing our own ideas
27 A: and then coming together
28 A: and um collaborating.

29 A: Okay,
30 A: we're gonna get to know each other a little bit.

31 A: So um,
32 A: what we're gonna do is start off with um let's start off with Amina.
33 A: Um Alima,
34 B: Alima.
35 A: sorry,
36 A: Alima.
37 A: Um we're gonna do a little tool training,
38 A: so we are gonna work with that whiteboard behind you.
39 A: Um introduce yourself,
40 A: um say one thing about yourself
41 A: and then draw your favourite animal
42 A: and tell us about it.
43 B: Okay.
44 B: Um I don't know which one of these I have to bring with me.
45 A: Probably both.

46 B: Right, so,
47 B: I'm supposed to draw my favourite animal.
48 B: I have no drawing skills whatsoever.

49 B: But uh let's see, introduce myself.
50 B: My name is Alima [Anonymized].
51 B: Um I'm from the state of [Anonymized] in the US.
52 B: I'm doing nationalism studies,
53 B: blah, blah, blah,
54 B: and I have no artistic talents.
55 . . .

[Transcript from *AMI Corpus*]

Text segmentation

As noted, **text segmentation** aims to make this high-level **linear structure** more explicit.

Why bother?

- Information can be **found** more effectively, which benefits tasks such as IR, IE, and QA;
- The properties of each type of segment can allow better summaries to be produced;
- One can develop **more accurate** segment-specific models of text that capture properties shared by all segments of a given type, which can benefits tasks such as MT [Foster, Isabelle & Kuhn, 2010].

Text segmentation

Text segmentation can be considered **low-hanging fruit** because

- decisions can be based on **proxies**;
- a **less than perfect** solution is acceptable, since even people produce only roughly similar segmentations.

Text segmentation

Proxies used in segmentation include:

- taking a segment to be a **bag** and/or **string** of tokens (words or word stems);
- using properties of bags or strings as evidence for segmentation decisions;
- using lexical or phrasal cues as additional evidence of the start or end of a segment.

Text segmentation

Fiction (*BNC*)	News (*WSJ*)	Parliament (*Hansard*)
Yes	In New York	To ask the
No	For the nine	The Prime Minister
What do you	In composite trading	My hon Friend
Oh	In early trading	Mr Speaker
What are you	In addition to	The hon Gentlemen
Of course	At the same	Order
Ah	One of the	Interruption
What's the	The White House	Does my right hon

[Sporleder & Lapata, 2006]

Text segmentation

Not all text segmentation is low-hanging fruit:

- hierarchical text segmentation;
- segmentation of texts whose high-level structure mirrors the speaker's own communicative intentions (**intentional structure**);
- segmentation of narrative text.

Nevertheless, enough is low-hanging for it to be a practical enterprise.

See [Purver, 2011] for more on topic-based segmentation, and [Webber et al, 2012] for more on genre-based segmentation.

Coherence relation recognition

Texts also have a **low-level structure** based on **coherence relations** between sentences and/or clauses.

Coherence relation recognition aims to identify **what** is connected and **how**.

Sometimes, the connection is **explicitly** marked:

- **inter-sententially**, by coordinating conjunctions or discourse adverbials, *inter alia*,
- **intra-sententially**, by coordinating or subordinating conjunctions, discourse adverbials, coordinators, *inter alia*

Sometimes, it is conveyed **implicitly**, via **adjacency**.

What in CRR are low-hanging fruit?

Coherence relation recognition

To answer this, need to understand the two main approaches to recognizing coherence relations:

- **text-centric** approach;
- **relation-centric** approach.

Coherence relation recognition

Text-centric approaches:

1. Divide a text into a sequence of adjacent discourse units;
2. Identify whether a relation holds between a pair of adjacent units and if so, what sense it conveys;
3. Add the result in as a derived discourse unit;
4. Continue until a tree structure of discourse units covers the text.

This is the approach taken in Rhetorical Structure Theory [Mann and Thompson, 1988] and automated approaches based on RST [Marcu, 2000; Sagae, 2009; Soricut & Marcu, 2003; Subba *et al*, 2006].

Coherence relation recognition

Relation-centric approaches:

1. Identify elements that **could** signal a coherence relation in a text and then check whether they actually do so.
2. Identify what each element relates (its **arguments**);
3. Identifying what sense it conveys.

This is the approach taken in the Penn Discourse TreeBank [Prasad et al., 2008] and similar discourse resources being developed for other languages (Arabic, Chinese, Italian, Turkish) and genres (journal papers in biomedicine, conversations).

Coherence relation recognition

Relation-centric approaches admit low-hanging fruit, since they can concentrate on frequent, easy-to-identify coherence relations.

This takes advantage of the **Zipfian distribution** of explicit discourse connectives.

Relation-centric approaches can also provide a partial solutions to coherence relation recognition by:

- Identifying an argument only in terms of its **head** [Wellner & Pustejovsky, 2007] or its **matrix sentence** [Prasad, Joshi & Webber, 2010];
- Identifying the sense of a relation only in terms of its high-level sense class [Pitler & Nenkova, 2009].

Coherence relation recognition

(10) Men have a tragic genetic flaw. As a result, they cannot see dirt until there is enough of it to support agriculture.
[Paraphrasing Dave Barry, The Miami Herald - Nov. 23, 2003]

Coherence relation recognition

(11) Men have a tragic genetic flaw. **As a result**, they cannot see dirt **until** there is enough of it to support agriculture.

Coherence relation recognition

(12) **Men have a tragic genetic flaw**. **As a result** [CONTINGENCY.RESULT], **they cannot see dirt until there is enough of it to support agriculture**.

(13) Men have a tragic genetic flaw. **As a result, they cannot see dirt until** [TEMPORAL.PRECEDENCE] **there is enough of it to support agriculture**.

Conclusion

- Research in NLP and LT starts by targetting low-hanging fruit made possible by
  - Zipfian distributions,
  - the availability of simpler (task-specific) proxies,
  - the acceptability of approximate solutions,
  - high-value recall.
- To understand distributions, it helps to have annotated corpora, which also allow us to test possible solutions.
- Once the low-hanging fruit is picked, one can go on to solve the challenging and often very informative problems raised by the long tail.

References

- George Foster, Pierre Isabelle & Roland Kuhn (2010). Translating structured documents. *Proceedings of AMTA*.
- William Mann & Sandra Thompson (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, **8(3)**, 243–281.
- Daniel Marcu (2000). The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, **26**, 395–448.
- Emily Pitler and Ani Nenkova (2009). Using Syntax to Disambiguate Explicit Discourse Connectives in Text. Proc. 47th Meeting of the Assoc. for Computational Linguistics and the 4th Int'l Joint Conf. on Natural Language Processing. Singapore.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber (2008). The Penn Discourse TreeBank 2.0. *Proc. 6th LREC*, Valletta, Malta.
- Rashmi Prasad, Aravind Joshi and Bonnie Webber (2010). Exploiting Scope for Shallow Discourse Parsing. *Proc. 7th Int'l Conference on Language Resources and Evaluation* (LREC 2010).

References

- Matthew Purver (2011). Topic Segmentation. In Gokhan Tur and Renato de Mori (eds.), *Spoken Language Understanding*, Wiley, 2011.
- Kenji Sagae (2009). Analysis of Discourse Structure with Syntactic Dependencies and Data-Driven Shift-Reduce Parsing. *In Proceedings of IWPT 2009*.
- Radu Soricut & Daniel Marcu (2003). Sentence Level Discourse Parsing using Syntactic and Lexical Information. *Proceedings of HLT/NAACL*.
- Caroline Sporleder and Mirella Lapata (2006). Broad coverage paragraph segmentation across languages and domains. *ACM Trans. Speech and Language Processing* 3(2), pp. 1–35.
- Rajen Subba, Barbara Di Eugenio and Su Nam Kim (2006). Discourse Parsing: Learning FOL Rules based on Rich Verb Semantic Representations to automatically label Rhetorical Relations. *In Proc. EACL Workshop on Learning Structured Information in Natural Language Applications*.
- Bonnie Webber, Markus Egg and Valia Kordoni (2012). Discourse Structure and Language Technology. *Natural Language Engineering*, 54 pages, doi:10.1017/S1351324911000337.

References

◦ Ben Wellner and James Pustejovsky (2007). Automatically Identifying the Arguments of Discourse Connectives. *Proc. Conf on Empirical Methods in Natural Language Processing*.