1. Introducing coreference

Consider the following narrative:

Violent strikes rocked Happyland. A spokesman for the country's Department of Peace said they would meet with the strikers tomorrow. Another spokesman said that this was intended to demonstrate the country's commitment to resolving the dispute.

It contains something like 12 noun phrases

- But there are only 7 entities being referred to

So there must be some shared referents

2. Some terminology

**Referring expression**
A part of an utterance used to identify or introduce an entity
- into a discourse (in real language use)
- into a discourse *model* (in a theory or implementation)

**Referents**
are such entities
- in a *model*
- or (imagined to be) in the world

**Reference**
is the relation between a **referring expression** and a **referent**

**Coreference**
When more than one **referring expression** is used to refer to the same entity

**Anaphora**
Reference to, or depending on, a previously introduced entity

3. Definite reference

NP's in red above are **definite** referring expressions

- as are the country and the dispute

Their use presupposes the existence of a unique (and uniquely identifiable) referent

- If a hearer does not already know of such a referent
  - Or, we will sometimes say, have such a referent in their *discourse model*
- They will usually accept the assumption that there is one
  - Or, we might say, add one to their model
- provided it's not inconsistent to do so
- This is known as **accommodation**

**Referring expressions** can be embedded in other **referring expressions**

- A spokesman for the country's Department of Peace
- the country's Department of Peace

4. Anaphora

The country (twice), they, the strikers, another spokesman, this, the dispute are **anaphoric expressions**

- They rely on the previous discourse for their interpretation.
- So the country, the strikers and the dispute are both definite referring expressions and anaphoric expressions.

5. Indefinite Reference

Indefinite NPs usually introduce new referents to the discourse:

- A spokesman for the Department of Peace said he would meet with the strikers.

When an indefinite NP is in the scope of **propositional attitude** verbs

- e.g. want, need, worry about
There is an ambiguity about its referent
Willard wants a sloop
• This might be a specific sloop that the speaker and/or Willard has in mind
• or an arbitrary, as yet unidentified, sloop
  ◦ As Quine puts it, he is seeking “mere relief from slooplessness”

6. Coreference

Referring expressions with the same referent are said to corefer.
Indefinite NPs can set up referents for subsequent coreferential anaphoric expressions:

There are fairies\textsubscript{i} in my garden. The fairies\textsubscript{i}/They\textsubscript{i} are having a ball

Not all fairies - just the ones in my garden
Not all subsequent anaphoric expressions corefer with their antecedent:

There are fairies\textsubscript{i} in my garden\textsubscript{m}. Other fairies\textsubscript{j} live elsewhere\textsubscript{k}

“other fairies” \neq fairies other than fairies\textsubscript{i}
• i.e., the ones in my garden\textsubscript{m}

“elsewhere” \neq places other than my garden\textsubscript{m}

7. Coreference and Pronouns

Pronouns serve as anaphoric expressions when they rely on the previous discourse for their interpretation

Definite pronouns
He, she, it, they etc.

Indefinite pronouns
One, some, elsewhere, other etc.

Some pronouns have other roles as well:
• periphrastic it: “It is raining”, “It is surprising that you ate a banana”
• generic they and one: “They’ll get you for that”, “One doesn’t do that sort of thing in public”

And some determiners have an anaphoric role:
• some, another, other etc.
• as in other fairies

The expression from the previous discourse used in interpreting a pronoun used anaphorically is called its antecedent.

A definite pronoun corefers with its antecedent.
The antecedent of an indefinite pronoun contributes in a more oblique way.

8. Reference resolution

Reference resolution is the process of determining the referent of a referring expression
• Whether by humans or machines

Context obviously plays a crucial role in reference resolution

Situational
The real-world surroundings (physical and temporal) for the discourse

Mental
The knowledge/beliefs of the participants

Discourse
What has been communicated so far

9. Discourse context—discourse model

For people we assume
• and in a computational system we construct
a discourse model
• That is, a set of “representations of the entities that have been referred to in the discourse and the relationships in which they participate” (J&M 21.3)

To produce and interpret referring expressions, a system must have methods for
• constructing a discourse model that evolves dynamically
• mapping between referring expressions and referents

In other words, for each referring expression, it must be able to determine when to
• Add a new entity the model to serve as the expression’s referent
  ◦ J&M call this evoking
• Find an existing entity ditto
  ◦ J&M call this accessing

10. Implementing reference resolution

Most approaches to implementing reference resolution distinguish two stages:
1. Filter the set of possible referents by appeal to linguistic constraints
2. Rank the resulting candidates based on some set of heuristics

11. Constraints on pronouns: Feature agreement

English pronouns agree with the number and/or gender of the referent of their antecedent.
Robin has a new car. It/*She/*They is red
Robin has a sister. *It/She/*They/*We is well-read
Robin has three cars. *It/*She/They/*We are all red

As well as the person (but case is determined locally):

• Robin and I/*me were late. *Me/*They/We/I missed the show
• Robin and I/*me were late. The usher wouldn't let *we/*I/us/me in

French pronouns agree with the number and gender of the form of their antecedent

• Voici une pomme. Je me demande si elle/*il/*elles est mûre [feminine form]
• Here’s an apple. I wonder if *she/it/*they is ripe [inanimate/neuter referent]

12. Constraints on pronouns: Syntax

All anaphors, including pronouns, rely on the previous text for all or part of their interpretation.

When the text is in the same sentence, pronominal coreference is subject to binding conditions

• John likes him vs. John likes himself
• John thinks Mary likes him/herself vs. *John thinks Mary likes himself
• Her brother admires Mary ⇒ Whose brother?
  • One reading, an example of cataphora

And, sometimes, to selectional restrictions based on the verb that governs it

John parked his car in the garage. He had driven it around for hours
  • It = the car, it ≠ garage
I picked up the book and sat in a chair. It broke
  • It = chair, it ≠ book

We will see how automated approaches to anaphora resolution exploit such constraints

13. Constraints aren’t enough

The kind of strong constraints we’ve just seen are not always enough to reduce the candidate set for resolution to a single entity

• John punched Bill. He broke his jaw/hand
• Barack hates his husband, but Hilary worked for/stays with him anyway

14. Heuristics for pronoun interpretation

Many different features influence how a listener will resolve a definite pronoun (i.e., what they will take to be its antecedent):

Recency
The most recently introduced entity is a better candidate
  • First Robin bought a phone, and then a tablet. Kim is always borrowing it

Grammatical role
- some grammatical roles (e.g. SUBJECT) are felt to be more salient than others (e.g., OBJECT)
  • Bill went to the pub with John. He bought the first round
  • “John” is more recent, but “Bill” is more salient.

15. More heuristics

Repeated mention
A repeatedly-mentioned entity is likely to be mentioned again

John needed portable web access for his new job. He decided he wanted something classy. Bill went to the Apple store with him. He bought an iPad.
  • “Bill” is the previous subject, but “John”’s repeated mentions tips the balance.

Parallelism
Parallel syntactic constructs can create an expectation of coreference in parallel positions
  • Susan went with Alice to the cinema. Carol went with her to the pub

16. Heuristics, concluded

Verb semantics
A verb may serve to foreground one of its argument positions for subsequent reference because of its semantics

John criticised Bill after he broke his promise vs. John telephoned Bill after he broke his promise

Louise apologised to/praised Sandra because she ...

World knowledge
At the end of the day, sometimes only one reading makes sense

• The city council denied the demonstrators a permit because they feared violence vs. The city council denied the demonstrators a permit because they advocated violence

17. Coreference: A more general case

Anaphoric pronoun resolution is a specific instance of the more general problem of coreference resolution

• Definite expressions other than pronouns are also candidates for reference resolution

Some of the heuristics enumerated above are relevant for the general case

• Particularly for more generic phrases such as “the guy” or “your man”

But other relations also come in to play
20. Using the model

Compute feature vectors for all possible referring expressions

For pronominal anaphora, we can just choose the most-positively scored (or largest positive vs. negative difference) antecedent

- Allowing for the possibility of no in-the-discourse coreferent at all

For definite referring expressions, choosing among available candidates versus not-in-the-discourse is a bit trickier

- And may require a separate model in its own right

21. Conclusion

There are usable "Off the shelf" coreference resolvers for English

- emPronoun, from Brown University
- BART, from Johns Hopkins
- Deterministic Coreference Resolution System, from Stanford
  - Which we'll use in the next assignment

There is still room for improvement in both coreference and anaphor resolution methods.

Knowing what expressions corefer and how other expressions relate to their antecedents can improve performance of Language Technology systems.
Lecture 25: Discourse, coherence, cohesion

Henry S. Thompson
(Based in part on slides by Johanna Moore and Bonnie Webber)
13 November 2014

1. "If we do not hang together then surely we must hang separately" (Benjamin Franklin)

Not just any collection of sentences makes a discourse.

- A proper discourse is coherent
- It makes sense as a unit
  - Possibly with sub-structure
- The linguistic cues to coherence are called cohesion

The difference?

- **Cohesion**: The (linguistic) clues that sentences belong to the same discourse
- **Coherence**: The underlying (semantic) way in which it makes sense that they belong together

2. Linking together

Cohesive discourse often uses lexical chains

- That is, sets of the same or related words that appear in consecutive sentences

Longer texts usually contain several discourse segments

- Sub-topics within the overall coherence of the discourse

Intuition: When the topic shifts, different words will be used

- We can try to detect this automatically

*But*, the presence of cohesion does not guarantee coherence

3. Identifying sub-topics/segmenting discourse

The goal is to delimit coherent sub-sequences of sentences

By division

- Look for cohesion discontinuities

By (generative) modelling

- Find the ‘best’ explanation

Relevant for

- Information retrieval
- Search more generally, in
  - lectures
  - news
  - meeting records
- Summarisation
  - Did we miss anything?
- Information extraction
  - Template filling
  - Question answering

4. Finding discontinuities: TextTiling

An unsupervised approach based on lexical chains

- Developed by Marti Hearst

Three steps:

1. Preprocess: tokenize, filter and partition
2. Score: pairwise cohesion
3. Locate: threshold discontinuities

5. TextTiling: Preprocessing

In order to focus on what is assumed to matter

- That is, content words

Moderately aggressive preprocessing is done:

- Segment at whitespace
- Down-case
- Throw out stop-words
- Reduce inflected/derived forms to their base
  - Also known as stemming
6. TextTiling: Scoring

Compute a score for the gap between each adjacent pair of token sequences, as follows:

1. Reduce blocks of k pseudo-sentences on either side of the gap to a bag of words:
   - That is, a vector of counts
   - With one position for every 'word' in the whole text
2. Compute the normalised dot product of the two vectors:
   - The cosine distance
3. Smooth the resulting score sequence by averaging the scores in a symmetrical window of width s around each gap

7. TextTiling: Locate

We're looking for discontinuities:
- Where the score drops
- Indicating a lack of cohesion between two blocks

That is, something like this:

\[ y_{i+1} \quad y_i \quad y_{i-1} \]

The depth score at each gap is then given by \((y_i - y_{i-1}) + (y_{i+1} - y_i)\)

Larger depth scores correspond to deeper 'valleys'

Scores larger than some threshold are taken to mark topic boundaries:
- Hearst evaluated several possible threshold values
- Based on the mean and standard deviation of all the depth scores in the document

Liberal

\[ s - \sigma \]

Conservative

\[ s - \frac{\sigma}{2} \]

8. Evaluating segmentation

How well does TextTiling work?
- Here's an illustration from an early Hearst paper

\[ y_i \]


The curve is smoothed depth score, the vertical bars are consensus topic boundaries from human readers.
- How can we quantify this?

Just classifying every possibly boundary as correct (Y+Y or N+N) vs. incorrect (Y+N or N+Y) doesn't work:
- Segment boundaries are relatively rare
  - So N+N is very common
  - The "block of wood" can do very well by always saying "no"

Counting just Y+Y seems too strict:
- Missing by one or two positions should get some credit

9. Evaluation, cont'd

The WindowDiff metric, which counts only misses (Y+N or N+Y) within a window attempts to address this.

Specifically, to compare boundaries in a gold standard reference (Ref) with those in a hypothesis (Hyp):
- Slide a window of size k over Hyp and Ref
- Compare the number of boundaries within the window at each possible position i in Ref (r) with those in Hyp (h):
  - That is, \(| r_i - h_i |\)
  - Count 0 if the result is 0 (correct)
  - Count 1 if the result is > 0 (incorrect)

Based on Figure 21.2 from Jurafsky and Martin 2009.

\[ 0 \] is the best result.
- No misses
10. Machine learning?

More recently, (semi-)supervised machine learning approaches to uncovering topic structure have been explored. Over-simplifying, you can think of the problem as similar to POS-tagging. So you can even use Hidden Markov Models to learn and label:

- There are transitions between topics
- And each topic is characterised by an output probability distribution

But now the distribution governs the whole space of (substantive) lexical choice within a topic:

- Modelling not just one word choice
- but the whole bag of words


11. Topic is not the only divider

Topic/sub-topic is not the only structuring principle we find in discourse:

- Different genres may mean different kinds of structure

Some common patterns, by genre:

**Expository**
- Topic/sub-topic

**Task-oriented**
- Function/precondition

**Narrative**
- Cause/effect, sequence/sub-sequence, state/event

But note that some of this is not necessarily universal:

- Different scholarly communities may have different structural conventions
- Different cultures have different narrative conventions
Different scholarly communities may have different structural conventions
Different culturals have different narrative conventions
Cohesion sometimes manifests itself differently for different genres

3. Functional Segmentation

Texts within a given genre
- News reports
- Scientific papers
- Legal judgements
- Laws

generally share a similar structure, independent of topic
- sports, politics, disasters
- molecular biology, radio astronomy, cognitive psychology

That is, their structure
- reflects the function played by their parts
- in a conventionalised structure

4. Example: news stories

The conventional structure is so 'obvious' that you hardly notice it
- Known as the inverted pyramid

In decreasing order of importance
- Headline
- Lead paragraph
  - Who, what, when, where, maybe why and how
- Body paragraphs, more on why and how
- Tail, the least important
  - And available for cutting if space requires it

5. Example: Scientific journal papers

In particular, experimental reports
- Your paper will not be published in a leading e.g. psychology research journal if it doesn't look like this

Highly conventionalised

Front matter
Title, Abstract

Body
(or, mnemonically, IMRAD
- Introduction (or Objective), including background
- Methods
- Results
- Discussion

Back matter
Acknowledgements, References
Although the major divisions (IMRAD) will usually be typographically distinct and of explicitly labelled
- Less immediately distinctive, more equivocal, cues give evidence for finer grained internal structure

6. Theories of discourse structure

Early discourse resources were task-oriented
- For example, an engineering explaining to an apprentice how to repair a pump
And the structure of task-oriented discourse often mirrored the structure of the task

Pre-computational theories had focussed on narrative structures
- Story grammars, so-called, basically taxonomic and flat

These gave way to structurally rich generative models
- Grosz and Sidner's Discourse Theory
- Mann and Thompson's Rhetorical Structure Theory (RST)
  - Not me, Sandra Thompson

Both were expressed in terms of coherence relations
- Also sometimes called discourse relations
- Between the interpretation of sentences/utterances
  - After some amount of abstraction

Still depending on observable phenomena (cohesion) to detect/identify them

7. Grosz and Sidner’s Discourse Theory (GSDT)

GSDT approaches the hierarchical nature of discourse at three levels
- Linguistic structure
  - What is actually said/written
- Intentional structure
  - Speaker’s goals and purposes
  - Organised into a relational structure
  - I.e. this is where discourse relations come in
- Attentional structure
  - Speaker’s focus of attention

8. GSDT: Intentional structure

There is an overall discourse purpose (DP)
- The basic purpose of the whole discourse

A discourse consists of discourse segments (DS)
- Each segment has one or more discourse segment purposes (DSP)
  - How a segment contributes to the DP
As well as **segment relations**

- **satisfaction-precedence** DSP1 must be satisfied before DSP2
- **dominance** DSP1 dominates DSP2 if fulfilling DSP2 constitutes part of fulfilling DSP1

9. Attentional state

Attention is represented in terms of a **focus stack**

- A stack of **focus spaces**
  - Each containing objects, properties and relations salient during its corresponding DS
  - As well as its DSP
- That is, content plus purpose

The discourse focus is always on the focus space at the top of the stack

State changes are modeled by transition rules controlling the addition/deletion of focus spaces

- Information at lower levels may or may not be available at higher levels

Focus spaces are pushed onto the stack when a new DS is detected

- And popped when they are completed

10. Discourse structure influences coreference

Consider our earlier example:

- **Welcome to word processing:**
  - [push]
    - That's using a computer to type letters and reports
    - Make a typo?
      - [push]
        - No problem
        - Just back up, type over the mistake, and it's gone
      - [pop]
    - And, it eliminates retyping

Or this

11. Detecting subtopics/Identifying discourse relations

There are a variety of cues that make these structures easier to recognise

- **implicit** lexical chains, tense and aspect
- **explicit** cue phrases
- **conjunctions** 'because', 'but'
- **conjunctive adverbials** 'nevertheless', 'instead'
- **temporal adverbials** 'then', 'afterwards'
- **suprasegmental** intonational variation

Different cues, different actions

- push
- pop
- chain
  - That is, pop then push

12. Learning cues

The Penn Discourse TreeBank provides a resource for supervised learning of cues

- Annotated with connectives, their arguments, and the senses they convey
- Approximately 18,000 explicit connectives and 16,000 implicit ones
As with many other linguistic phenomena, the distribution of connectives follows Zipf’s law

- but 3308
- and 3000
- if 1223
- because 858
- while 781
- however 465
- therefore 26
- otherwise 24
- as soon as 20
- accordingly 5
- if and when 3
- conversely 2

The corpus isn’t large enough to have all the connectives we might expect

13. Discourse relations

What kinds of things are we looking for?
- Kinds of coherence

How do discourse segments “hang together”?

Essentially an abductive question
- Abduction is reasoning to the best explanation

Which often is explanation

Compare

John hid Bill’s car keys. He was drunk
John hid Bill’s car keys. He likes spinach

Here’s a list from Mann & Thompson:
- Circumstance
- Solutionhood
- Elaboration
- Background
- Enablement and Motivation
  - Enablement
  - Motivation
- Evidence and Justify
  - Evidence
  - Justify
- Relations of Cause
  - Volitional Cause
  - Non-Volitional Cause
  - Volitional Result
  - Non-Volitional Result
- Purpose
- Antithesis and Concession
  - Antithesis
  - Concession
- Condition and Otherwise
  - Condition

14. RST, cont’d

Relations defined by constraints on the nucleus and satellite

- With respect not only to N and S
- But also the writer (W) and the reader (R)

For example, the Evidence relation

<table>
<thead>
<tr>
<th>Relation Name</th>
<th>Constraints on N</th>
<th>Constraints on S</th>
<th>Constraints on N+S</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence</td>
<td>R might not believe N to a degree satisfactory to W</td>
<td>R believes S or will find it credible</td>
<td>R’s comprehending S increases R’s belief of N</td>
<td>R’s belief of N is increased</td>
</tr>
</tbody>
</table>

Relations may be either symmetric (e.g. Contrast) or asymmetric (e.g. Purpose)

- There’s a set of graphical conventions for diagramming relations and their overall pattern in a discourse

![Figure 21.4](image-url) A discourse tree for the Scientific American text in (21.23), from Marcus (2000a). Note that asymmetric relations are represented with a curved arrow from the satellite to the nucleus.
15. Contrasting approaches to coreference in discourse

A motivating antecedent to Grosz’s discourse work was Hobbs’s 1978 pronoun resolver

- Purely syntactic
- Gender, person, number checks
- Carefully-crafted search order
  - Up from the pronoun
  - Across left-to-right, breadth-first, for each S or NP as you go up
  - Back to the previous sentences


Grosz (at SRI) and Sidner (at MIT) moved from a flat approach to one dependent on hierarchical discourse structure

- Called Centering theory
- Key roles for two kinds of center:
  - One backward-looking
  - Multiple forward-looking
- And rules associated with push, pop etc. for updating these

See J&M 21.6.1 and 21.6.2 for detailed expositions of these two approaches

1. Words, and other words

A brief introduction to machine translation

**Machine Translation** covers a wide range of goals

- **From FAHQUMT**
  - Fully Automatic High Quality Unrestricted MT
- **To MAHT**
  - Machine-Assisted Human Translation
- FAHQUMT remains a dream: but hope springs eternal
- MAHT is big business, but not of much theoretical interest

The contrast between hyped-up promises of success and poor actual performance led to

- The ALPAC report (1966)
- Found that many years of research had failed to meet expectations
  - USA has no shortage of translators
  - Fully automatic MT doesn’t really work, quality hasn’t improved much
  - It isn’t clear if it will ever work

“The Committee indeed believes that it is wise to press forward undaunted, in the name of science, but that the motive for doing so cannot sensibly be any foreseeable improvement in practical translation. Perhaps our attitude might be different if there were some pressing need for machine translation, but we find none.”

- The end of substantial funding for MT in the US for nearly 20 years

2. From knowledge-rich to machine-learning

MT has followed the same trajectory as many other aspects of speech and language technology

- Historically, MT systems were based on one or more levels of linguistic analysis
  - the Vauquois triangle
  - The largest and most heavily used MT system in the world worked like this until very recently
    - SYSTRAN, used by the EU to help with its translation load of over 2 million pages a year
- But most MT work today is based on one form or another of noisy channel decoding
  - With language and channel models being learned from corpora
3. Before his time: Warren Weaver

Stimulated by the success of the codebreakers at Bletchley Park (including Alan Turing), Weaver had an surprisingly prescient idea:

...(3) knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography...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.” Have you ever thought about this? As a linguist and expert on computers, do you think it is worth thinking about?

from a letter from Warren Weaver to Norbert Wiener, dated April 30, 1947

4. A very noisy channel

Applying the noisy channel model to translation requires us to stand normal terminology on its head

- Usually we talk about source and target languages
- For example, when translating Братья Карамазовы into English
  - Russian is the source
  - English is the target

But from the perspective of the noisy channel model

- The source is English
- The channel distorts this into what we see or hear, that is, Russian
- Which we have to decode
  - to get to the source
  - which is the target
  - :-(

5. Priors and likelihood for MT

Remember the basic story (using e for English and r for Russian):

<table>
<thead>
<tr>
<th>Prior</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P[e_1^n] )</td>
<td>( P[r_j</td>
</tr>
</tbody>
</table>

\[
\arg\max_{e_1^n} P[r_1^n | e_1^n P[e_1^n]]
\]

The prior is just our old friend, some form of language model

- for English!

But the channel model needs to be articulated a bit for translation, in several ways

- The source and target need not have the same number of words
- And the mapping part, on even a fairly simple view, has to do two things:
  - Not just translate the words
  - But re-order them as well

So we need a channel model that takes all of these into account

6. Translation modeling for MT

J&M Chapter 25 takes you through a step-by-step motivation for the first successful attempt at doing things this way

- By the IBM team, working from French into English
- Using Canadian Hansard for training and testing
- And a variety of HMM-based decoding methods

All their approaches start with a formal notion of alignment

- A (possibly one-to-many) mapping from source word position to position in the observation sequence
- So, for our trivial Russian example, this would be 1 3 2
- Because the second and third words exchange positions between English and Russian

Я вас любил
I loved you

7. Translation modelling, cont'd

Their simplest model then has three conceptual steps:

1. Choose a length for the Russian, given the length of the English
   - Remember, from the model’s perspective we are generating the Russian observation, starting from the English source
   - Think of a POS-tagging HMM
   - Which ‘generates’ English words (the observations) from a sequence of POS tags (the source)
2. Choose an alignment from the words in the source (English) to the words in the observations (Russian)
3. For each position in the Russian, choose a translation of the English word which aligns to it

Following simplifying assumptions of the usual Markov nature, we end up with

\[
P(r_1^n, e_1^n | e_1^n P[e_1^n]) = \prod_{j=1}^{n} P(a_j | a_{j-1}, j) P(r_j | a_j)
\]

Where

- \( i \) and \( j \) are the (underlying) length of the English and the (generated) length of the Russian, respectively
8. Contemporary MT

We've barely scratched the surface

But state-of-the-art MT systems today all derive from essentially this starting point

- Including Google Translate
- Which does however think that “я вас любил” should be translated as “I loved you more”
- To be fair, that's in the context of the whole Pushkin poem from which the phrase is extracted

9. Getting started: The role of data

Broadly speaking, we have two models to learn:

**Language model**
- We've seen this already,
  - Target language data
  - I.e. monolingual
  - Lots of it
  - Such as Google's Gigaword corpus

**Channel/translation model**
- For word-word alignment and word/phrase translation
  - Bilingual
  - Harder to get lots

These will vary in difficulty given the form of the raw data

- And the language involved

But for the translation model, with respect to the bilingual data, we need more

- We need to align the two versions at the paragraph and sentence level
- Sentence level is not always 1-to-1

10. Getting started: segmentation and sentence alignment

Just as with other corpora, we need to pre-process the raw materials

- Normalise markup
- Check for and correct character encoding problems
- Segment and normalise
  - tokens
  - morphemes?
  - sentences
  - paragraphs
  - down-case at beginning of sentences, maybe
  - tag

11. Sentence alignment details: Gale and Church (1993)

Assumptions:

- We start with two documents
  - In source and target languages
  - Translations of one another
- Sentence order is rarely if ever changed
- If paragraphs exist, they are already aligned

Paragraph by paragraph, the algorithm matches source sentences to zero, one or two target sentences

- Sentence may be deleted by translator
- Sentence may be split into two by translator
- In either direction
  - We don't actually always know which was the original

12. Gale and Church, cont'd

Start with some empirical observations:

What does a hand-aligned corpus tell us about sentence alignment?
That gives G&C the basis for a maximum likelihood estimate of $P(\text{match})$.

Where by match is meant a particular alignment choice.

What about relative length?

- If we suppose lengths (in characters) are normally distributed around equality.
- With standard deviation estimated from the same hand-aligned corpus.
- We get this picture when we plot the actual z-scored ratio.

G&C call this $\delta$:

$$\delta = \frac{l_2 - l_1}{\sqrt{l_1 s_2^2}}$$

- With a little more work, this gives G&C an estimate of $P(\delta | \text{match})$.

13. Gale and Church, concluded

(You may see where this is headed by now)

Given a candidate local alignment (a match), we can now compute

- $P(\delta | \text{match})$, the probability of that match having that value for $\delta$.
- $P(\text{match})$, the probability of that kind of match (2->1, 1->0, etc.)