Discourse and SMT

Bonnie Webber

School of Informatics
University of Edinburgh
bonnie@inf.ed.ac.uk

September 10, 2014

What is discourse?

Definition 1 (J&M, p. 715)

> *Discourse involves a collocated, structured, coherent group of sentences.*

Definition 2:

> *Discourse involves a sequence of* $\geq$*2 clauses whose content is related either through their meaning or through what the speaker is trying to do with them.*

Discourse and sentence sequences

Evidence for discourse simply involving $\geq$2 clauses in a single sentence:

(1) If they're drunk
    and they're meant to be on parade
    and you go to their room
    and they're lying in a pool of piss,
  then you lock them up for a day.
  [The Independent, 17 June 1997]

Chinese sentences are often small discourses: Over **25%** of Chinese sentences correspond to $>$1 sentence in their English reference text [Li, Carpuat & Nenkova, 2014].

Overview

Here I will address ways in which discourse shapes text, which makes things either more difficult or easier for SMT.

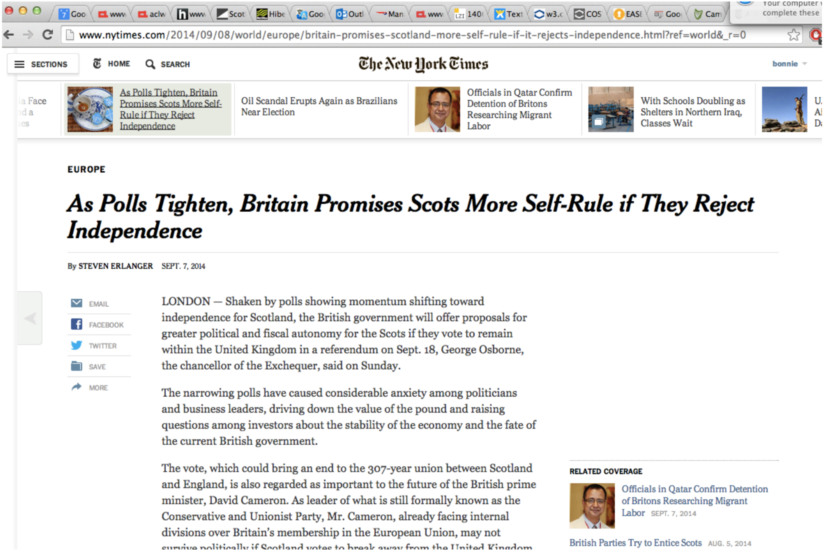
- A text is rarely a simple sequence of sentences: Rather, it is organized into a **macro-level structure** common to its genre.
- Structure is characterized by both **topic** and **function**. Both shape the patterns of words found within them. Translation into any target language should be **consistent with** this.
- Discourse sets up a **context** that speakers use to convey information more efficiently, using language-specific, **context-dependent devices** such as pronouns and other anaphoric devices.

Overview

- A text is not simply a sequence of unrelated clauses and/or sentences: Rather, there are semantic and/or pragmatic **discourse relations** between them, which may be signalled explicitly or left to inference.

- By and large, discourse features have a **locality** that challenges the standard sentence-oriented paradigm of SMT. Current research in discourse and MT aims to overcome this challenge efficiently (e.g., through **caching** and **document-level MT**).

Macro-structures in Discourse: Examples



www.nytimes.com/2014/09/08/world/europe/britain-promises-scotland-more-self-rule-if-it-rejects-independence.html?ref=world&_r=0

As Polls Tighten, Britain Promises Scots More Self-Rule if They Reject Independence

Oil Scandal Erupts Again as Brazilians Near Election

Officials in Qatar Confirm Detention of Britons Researching Migrant Labor

With Schools Doubling as Shelters in Northern Iraq, Classes Wait

EUROPE

As Polls Tighten, Britain Promises Scots More Self-Rule if They Reject Independence

By STEVEN ERLANGER SEPT. 7, 2014

LONDON — Shaken by polls showing momentum shifting toward independence for Scotland, the British government will offer proposals for greater political and fiscal autonomy for the Scots if they vote to remain within the United Kingdom in a referendum on Sept. 18, George Osborne, the chancellor of the Exchequer, said on Sunday.

The narrowing polls have caused considerable anxiety among politicians and business leaders, driving down the value of the pound and raising questions among investors about the stability of the economy and the fate of the current British government.

The vote, which could bring an end to the 307-year union between Scotland and England, is also regarded as important to the future of the British prime minister, David Cameron. As leader of what is still formally known as the Conservative and Unionist Party, Mr. Cameron, already facing internal divisions over Britain's membership in the European Union, may not survive politically if Scotland votes to break away from the United Kingdom

RELATED COVERAGE

Officials in Qatar Confirm Detention of Britons Researching Migrant Labor SEPT. 7, 2014

British Parties Try to Entice Scots AUG. 5, 2014

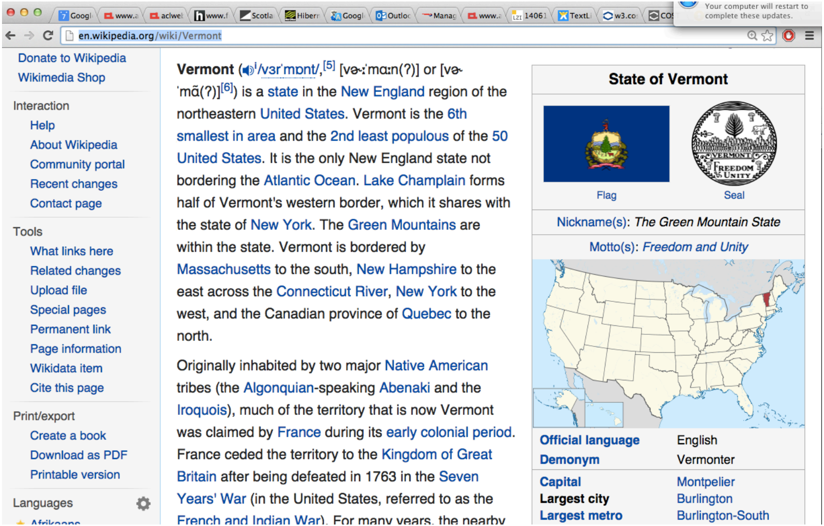
Macro-structures in Discourse: Examples

News reports are written with an **inverted pyramid** structure:

- Headline
- Lead 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, . . .

This is why the first (ie, lead) paragraph is usually the best *extractive summary* of a news report.

Macro-structures in Discourse: Examples



Vermont (/vɜrˈmɒnt/,[5] [vɚːˈmɑːn(ʔ)] or [vɚˈmɑ̃(ʔ)][6]) is a state in the New England region of the northeastern United States. Vermont is the 6th smallest in area and the 2nd least populous of the 50 United States. It is the only New England state not bordering the Atlantic Ocean. Lake Champlain forms half of Vermont's western border, which it shares with the state of New York. The Green Mountains are within the state. Vermont is bordered by Massachusetts to the south, New Hampshire to the east across the Connecticut River, New York to the west, and the Canadian province of Quebec to the north.

Originally inhabited by two major Native American tribes (the Algonquian-speaking Abenaki and the Iroquois), much of the territory that is now Vermont was claimed by France during its early colonial period. France ceded the territory to the Kingdom of Great Britain after being defeated in 1763 in the Seven Years' War (in the United States, referred to as the French and Indian War). For many years, the nearby

State of Vermont

Nickname(s): *The Green Mountain State*

Motto(s): *Freedom and Unity*

Official language	English
Demonym	Vermonter
Capital	Montpelier
Largest city	Burlington
Largest metro	Burlington-South

Macro-structures in Discourse: Examples

Expository text is written in **topically coherent segments**, whose order may become conventionalized over time:

	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	...	...	...

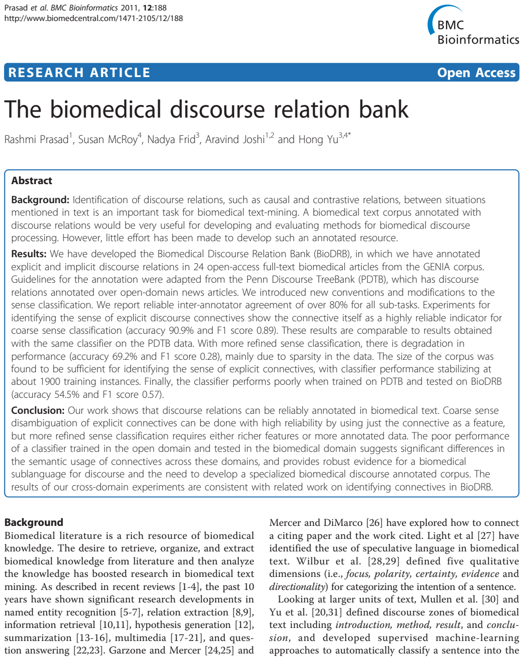
Wikipedia articles about US states

Macro-structures in Discourse: Examples

Scientific papers are structured into functionally-specific parts:

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

Their **structured abstracts** have a similar structure.



Prasad *et al. BMC Bioinformatics* 2011, **12**:188
http://www.biomedcentral.com/1471-2105/12/188

BMC Bioinformatics

RESEARCH ARTICLE Open Access

The biomedical discourse relation bank

Rashmi Prasad[1], Susan McRoy[4], Nadya Frid[3], Aravind Joshi[1,2] and Hong Yu[3,4*]

Abstract

Background: Identification of discourse relations, such as causal and contrastive relations, between situations mentioned in text is an important task for biomedical text-mining. A biomedical text corpus annotated with discourse relations would be very useful for developing and evaluating methods for biomedical discourse processing. However, little effort has been made to develop such an annotated resource.

Results: We have developed the Biomedical Discourse Relation Bank (BioDRB), in which we have annotated explicit and implicit discourse relations in 24 open-access full-text biomedical articles from the GENIA corpus. Guidelines for the annotation were adapted from the Penn Discourse TreeBank (PDTB), which has discourse relations annotated over open-domain news articles. We introduced new conventions and modifications to the sense classification. We report reliable inter-annotator agreement of over 80% for all sub-tasks. Experiments for identifying the sense of explicit discourse connectives show the connective itself as a highly reliable indicator for coarse sense classification (accuracy 90.9% and F1 score 0.89). These results are comparable to results obtained with the same classifier on the PDTB data. With more refined sense classification, there is degradation in performance (accuracy 69.2% and F1 score 0.28), mainly due to sparsity in the data. The size of the corpus was found to be sufficient for identifying the sense of explicit connectives, with classifier performance stabilizing at about 1900 training instances. Finally, the classifier performs poorly when trained on PDTB and tested on BioDRB (accuracy 54.5% and F1 score 0.57).

Conclusion: Our work shows that discourse relations can be reliably annotated in biomedical text. Coarse sense disambiguation of explicit connectives can be done with high reliability by using just the connective as a feature, but more refined sense classification requires either richer features or more annotated data. The poor performance of a classifier trained in the open domain and tested in the biomedical domain suggests significant differences in the semantic usage of connectives across these domains, and provides robust evidence for a biomedical sublanguage for discourse and the need to develop a specialized biomedical discourse annotated corpus. The results of our cross-domain experiments are consistent with related work on identifying connectives in BioDRB.

Background

Biomedical literature is a rich resource of biomedical knowledge. The desire to retrieve, organize, and extract biomedical knowledge from literature and then analyze the knowledge has boosted research in biomedical text mining. As described in recent reviews [1-4], the past 10 years have shown significant research developments in named entity recognition [5-7], relation extraction [8,9], information retrieval [10,11], hypothesis generation [12], summarization [13-16], multimedia [17-21], and question answering [22,23]. Garzone and Mercer [24,25] and Mercer and DiMarco [26] have explored how to connect a citing paper and the work cited. Light et al [27] have identified the use of speculative language in biomedical text. Wilbur et al. [28,29] defined five qualitative dimensions (i.e., *focus, polarity, certainty, evidence* and *directionality*) for categorizing the intention of a sentence.

Looking at larger units of text, Mullen et al. [30] and Yu et al. [20,31] defined discourse zones of biomedical text including *introduction, method, result*, and *conclusion*, and developed supervised machine-learning approaches to automatically classify a sentence into the

Using Discourse Macro-structures in SMT

Foster, Isabelle & Kuhn [2010] demonstrated the possibility of structured document translation, using the Hansard FR–EN corpus of transcripts of Canadian Parliamentary proceedings.

Their segments correspond to individuals speaking to particular issues in particular languages, since everyone uses language in their own way (*individual variation*, *stylistics*).

- Inducing and using Language Models for each combination of features (source language, speaker, activity, year),
- and then generalizing over certain combinations to reduce sparcity,
- produced modest, but statistically significant improvement in BLEU score, in both translation directions.

Using Discourse Macro-structures for SMT

Predictable sub-topic structure of expository text admits the possibility of sub-topic level **domain adaptation** [Louis & Webber, 2014].

This involves:

- separately inducing **segment-level subtopic models** for source and target languages from comparable domain-specific corpora,
- linking the S/T subtopic models using a dictionary,
- during translation, inferring the topic for each segment in the source text;
- loading a **topic cache** with words likely for next segment's topic,
- scoring each word in the cache to reflect its probability under the topic.

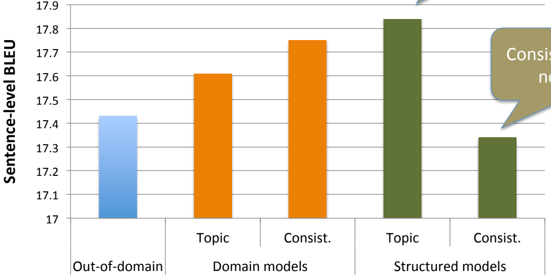
Using Discourse Macro-structures for SMT

For tuning and test sets, [Louis & Webber 2014] used translated articles from Wikipedia that were

- filtered via metadata to get **biographies** in French and English,
- of approximately the **same length** (so alignable) – 12 to 87 sentences each
- with **$\geq$3 section headings** (so can test structure hypothesis) – average=5
- Tuning: 13 pairs of articles; Test: 30 pairs

Graphics on next slides provided by Annie Louis

Using Discourse Macro-structures for SMT



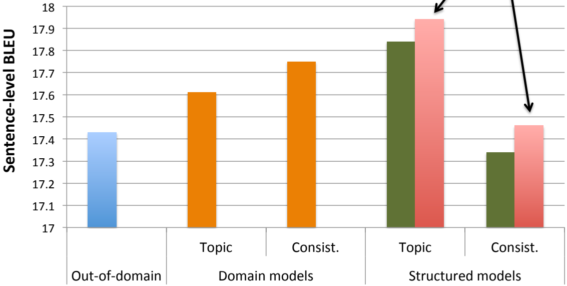
Struct-topic gives best performance

Consistency cache not useful

Using Discourse Macro-structures for SMT

- Approach performed best with ∼50 topics.
- Topic cache worked better for longer documents: Highest average gain in BLEU for documents with 30-49 sentences.
- Topic cache worked better for longer segments: Highest average gain in BLEU for segments with 11-17 sentences.

Using Discourse Macro-structures for SMT



Automatic segmentation gives better results

A good segmentation is important to performance of struct-models

Context-Dependent Devices: Anaphora

For anaphoric expressions, their form and meaning depend (in whole or in part) on the **context** established by the previous discourse.

We usually make the **simplifying assumption** that the **previous text** can serve as an **effective proxy** for context.

Languages have many different kinds of anaphoric expressions, including (in English):

- personal pronoun anaphors
- definite NP anaphors
- comparative anaphors (e.g., *Lawyers and Other Reptiles*)
- S and VP anaphors (dependent on previous events or actions)
- bridging anaphors (e.g., *bus* / *driver*)

Context-Dependent Devices: Anaphora

Accurate translation of context-dependent devices can be difficult for SMT.

For personal pronoun anaphors, their form must agree with features of their antecedent.

- The **U.S.A.**, claiming some success in **its** trade diplomacy, ... → inanimate sg. possessive
- **USA** tvrdí někteří úspěchu v **své** obchodní diplomacii ... → inanimate masculine pl. possessive

Achieving such agreement has been a focus of work in this area: Le Nagard & Koehn [2010]; Hardmeier & Federico [2010]; Novák [2011]; Guillou [2012].

We can see if this is the right way to think about and address the problem, by looking at pronouns in parallel corpora.

Context-Dependent Devices: Anaphora

PCEDT (Prague Czech-English Dependency TreeBank): Manually annotated parallel treebank (http://ufal.mff.cuni.cz/data/pcedt) aligned with the **Penn WSJ Corpus**.

PCEDT coreference annotation can be compared with **OntoNotes** coreference annotation of Penn WSJ Corpus (used in CoNNL-2011 Shared Task).

N.B. I don't know if anyone has looked at how coreference usage is similar or different in the two corpora.

Context-Dependent Devices: Anaphora

ParCor [Guillou et al, 2014] aims to:

- provide gold standard test sets for SMT;
- help identify systematic differences in pronoun use between languages;
- help build discourse-informed SMT systems.

ParCor annotates pronouns by **type** (e.g. anaphoric, pleonastic, event), with features appropriate to each type, including **antecedents** for anaphoric pronouns.

ParCor 1.0 comprises 19 English texts translated into German:

- 8 long documents from the EUBookshop publications
- 11 TED talks from the IWSLT'13 "2010" test set

Soon to be added are 2 parallel annotated TEDx talks (German source, English target).

Analysis of Sample Document from ParCor 1.0

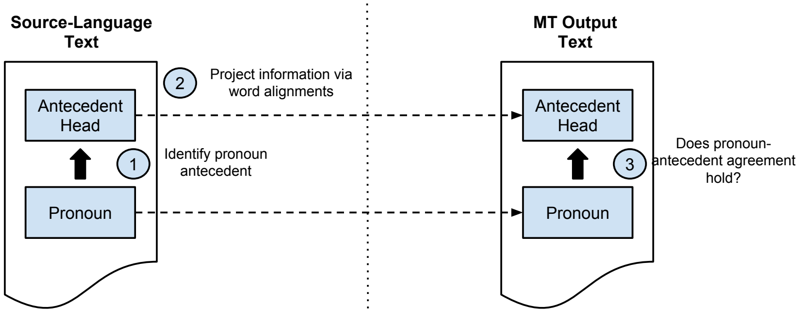
In 259 manually-translated sentences, with 444 source (English) pronouns and 564 target (German) pronouns, there were:

- 322 matches (pronoun pairs) by raw count
- **364 mismatches**

Type	Source (EN)	Target (DE)
Anaphoric	**35**	**103**
Pleonastic	**3**	**49**
Event	23	33
Addressee Reference	30	19
...	...	...

Personal pronouns in SMT

Annotation projection uses alignments and SL pronoun-antecedent links to identify pronouns for which agreement holds, as well as mismatches [Le Nagard & Koehn, 2010; Hardmeier & Federico, 2010; Guillou, 2012].


Source-Language Text
MT Output Text
Antecedent Head
Pronoun
1
Identify pronoun antecedent
2
Project information via word alignments
Antecedent Head
Pronoun
3
Does pronoun-antecedent agreement hold?

(Graphics provided by Liane Guillou)

Mismatch: Relative pronouns

English can signal a relative clause using a relative pronoun (**that**, **which**, **who**) or null (Φ), while German uses a personal pronoun.

Source: And this is a problem **that** a lot of gamers have.
Ref (TED): Und das ist ein Problem , **das** viele Spieler haben.
Google: Und das ist ein Problem , **dass** viele Spieler zu haben. χ

Source: The second factor is the services Φ we use.
Ref (TED): Der zweite Faktor sind die Dienste, **die** wir benutzen.
Google: Der zweite Faktor ist die Dienstleistungen, **die** wir verwenden. $\surd$

Pronouns and Evaluation

- BLEU is too coarse to assess the treatment of any linguistic phenomenon.
- Human evaluation is slow and expensive.
- MT output will generally have a pronoun if there is one in the source (though they too can be dropped).
- But a reference translation may use a pronoun construction for a pronoun-free source or a pronoun-free construction for a source pronoun.

Source: Also hat er vor den **Spielern** eine Kamera aufgebaut.
Ref (TED): So, he set up a camera in front of **gamers** while **they** were playing.
Google: So he set up a camera in front of the **players**.

Pronouns and Evaluation

One might try to assess some cases automatically, cf. ACT [Hajlaoui & Popescu-Belis, 2013]

- Check if pronoun in MT output matches one in reference.
- Ignore cases where pronoun in source and MT output but not in reference: May be OK but not fluent.
- Learn to ignore some expressions with a pronoun mismatch: **There is** → **Es gibt**.
- Check agreement of the rest via **annotation projection**

Manual evaluation still needed for pronouns in reference but not in MT output (and possibly not in the source):

- Languages differ in where they allow coreference to be expressed implicitly.
- Where they allow both antecedent–pronoun and pronoun-free constructions, may be able to use **paraphrase** to generate allowable alternatives.

Other Context-Dependent Constructions in SMT

In some cases, **how** a construction depends on the context may not matter. E.g.

- In German–English translation, we don't appear to need to know what an **event anaphor** refers to in order to translate it correctly.
- The same holds for **comparative anaphors** (e.g. *other countries* ↔ *andere Länder*)
- An NP may be definite because
  - it refers to an entity treated as **unique** (*the sun*), or
  - it is **anaphoric** (*a student who attended MT Marathon* → *the student*), or
  - it is a **bridging anphor** (*a bus to Povo* → *the driver*).

  Distinguishing them may be needed for correct translation.

Discourse Relations and SMT

- Quick overview of discourse relations
- Disambiguation 1: Recognizing a token as a discourse connective, for correct translation
- Diambiguation 2: Identifying the sense of a discourse connective, for correct translation
- Projecting constraints on the arguments to a discourse relation
- Final problem: Languages differ in how often they use explicit discourse connectives

Quick Overview of Discourse Relations

Discourse is more than its individual sentences. Clauses and sentences tend to relate to each other in terms of

- their topics
- the entities they refer to
- particular semantic and pragmatic relations taken to hold between them.

Readers (hearers) derive these semantic and/or pragmatic **relations** on the basis of:

- world knowledge
- general cognitive biases
- linguistic features, including explicit discourse connectives.

Quick Overview of Discourse Relations

With sufficient evidence, no explicit connective may be required to infer intended discourse relations:

(2) He suspected he shouldn't interrupt the speaker with a question. He should wait until the end of the talk.

⇒ no need for instead

But this doesn't mean an explicit connective isn't allowed:

(3) He suspected he shouldn't interrupt the speaker with a question. Instead he should wait until the end of the talk.

Or even two:

(4) He suspected he shouldn't interrupt the speaker with a question, but instead he should wait until the end of the talk.

Overview: Explicit Discourse Connectives

In many languages, the **discourse connectives** taken to explicitly signal discourse relations come from well-defined syntactic classes:

- **Subordinating conjunctions**: *because, although, when, if*, etc.
- **Coordinating conjunctions**: *and, but, so, nor, or* (and paired versions of the latter – *either..or, neither..nor*)
- **Prepositional phrases**: *as a result, on the one hand..on the other hand, insofar as, in comparison*, etc.
- **adverbs**: *then, however, instead, likewise, subsequently* etc.

Overview: Alternative Lexicalizations

But there are other types of evidence for discourse relations, called *alternative lexicalizations* or **AltLex** in the Penn Discourse TreeBank (PDTB) [Prasad et al, 2008, 2014] and the Prague Dependency TreeBank 3.0 [Rysova 2012].

(5) **The two companies each produce market pulp, containerboard and white paper**. **That means goods could be manufactured closer to customers, saving shipping costs**, he said. [wsj_0317]

(6) **The new structure would be similar to a recapitalization in which holders get a special dividend yet retain a controlling ownership interest**. **The difference is that current holders wouldn't retain majority ownership or control**. [wsj_1531]

This extends evidence for discourse relations beyond well-defined syntactic classes.

Disambiguation 1: Recognizing Tokens as DConns

Expressions that function as discourse connectives often have other roles as well, sometimes with a different PoS-tag, sometimes not.

Appropriate target translation can depend on role.

(7) Asbestos is harmful **once** it enters the lungs.
subordinating conjunction → *wenn*, *nachdem*

(8) Asbestos was **once** used in cigarette filters.
adverb → *einmal*, *einst*

Disambiguation 1: Recognizing Tokens as DConns

Pitler & Nenkova [2009] were able to distinguish discourse and non-discourse usage of tokens with an f-score of

- 75.33% based on the string alone
- 88.19% based on syntactic features of Gold Standard parse
- 92.28% using both
- 94.19% also using interactions between syntactic features.

For translation, one wants accuracy figures for tokens whose translation differs, depending on discourse vs. non-discourse role.

Disambiguation 2: Recognizing the Sense(s) of a DConn

Some discourse connectives are **ambiguous**: They can be used to convey more than one sense.

(9) Vicar Marshall admits to mixed feelings about this issue, since [**Cause.Reason**] he is both a vicar and an active bell-ringer himself. [wsj_0089]

(10) Mr. Bernstein, who succeeded Bennett Cerf, has been only the second president of Random House since [**Temporal.Succession**] it was founded in 1925. [wsj_0111]

As with other ambiguities, appropriate translation depends on sense.

Connective sense is **local**: Not one sense or one translation per discourse.

Disambiguation 2: Recognizing the Sense(s) of a DConn

For EN-FR translation, Meyer & Popescu-Belis [2012] have:

- induced sense classifiers from the Penn Discourse TreeBank (PDTB) for 13 sense-ambiguous EN discourse connectives whose proper translation into FR depends on their sense;
- automatically found tokens of those discourse connectives in their training, tuning and test corpora (cf. Disambiguation 1);
- used the sense classifiers to label those tokens;
- As a sanity check, because of **register** differences, manually labelled 5 sense-ambiguous EN connectives and their counterparts on the FR side (∼1200 tokens);
- experimented with ways of using labelled tokens in SMT;
- evaluated the results on both automatically and manually labelled connectives.

Disambiguation 2: Recognizing the Sense(s) of a DConn

Many different experiments reported in [Meyer & Popescu-Belis, 2012], including

- training on manual annotations (5 connectives) and testing on automatically labelled tokens of those connectives.
- training on automatic annotations (13 connectives) and testing on automatically labelled tokens of those connectives.

Two evaluations:

- BLEU score (0.5 – 1.5 point increase)
- manual assessment of better–same–worse choice of connective (% change)

	+	=	-
manual5/auto5	34%	46%	20%
auto13/auto13	16%	60%	24%

Using constraints on DRel arguments for SMT

Certain discourse relations can only hold if one or the other or both of their arguments have certain features.

The relation variously called SUBSTITUTION, REPLACEMENT, or CHOSEN ALTERNATIVE is a case in point: One arg is the source of the **replacing** alternative and the other, the source of the alternative that can be **replaced**. [Webber, 2013].

(11) **Alice could tell Ed everything in half an hour**, **but** instead **she drags her story out**. [TLS, 1 Jun 2012]

(12) **Also können sie nicht Teil einer breiteren Linie sein** und **müssen** stattdessen **Dinge aus ihrer eigenen Position heraus vorantreiben**.

Thus they are unable to be part of a wider authority and must instead drive matters forward from their own position.

Constraints on Argument Features

There are 118 tokens of explicit CHOSEN ALTERNATIVE in the PDTB and 171 implicits.

47 of the explicits (39.8%) and 116 of the implicits (67.8%) have a **negation** marker (e.g., *not*, *no*, *never*, *nothing*, *nobody*, etc.) in **Arg1**, the source of the replaced alternative.

(13) If the flex is worn, **do not use insulating tape to repair it**. Instead, **you should replace it . . .** .

(14) **There are no separate rafters in a flat roof**; instead, **the ceiling joists of the top story support the roofing**.

Constraints on Argument Features

While I don’t have counts, negation markers are found in **Arg1** of similar relations in German.

(15) Es reiche **nicht** aus, sich auf internationale Markenrechte zu verlassen, vielmehr sollten die Ausländer “alles, was irgendwie schützenswert ist, auch in China anmelden”, wie Wentzler sagt.

It is **not** *sufficient to rely on international trademark rights, rather foreigners should also register “everything that is in any way worthy of protection in China as well,” said Wentzler.*

(16) Oder sie wählen überhaupt **kein** Gerät und zahlen stattdessen eine Pauschalgebhr auf Grundlage der durchschnittlich von allen Einwohnern des Bundesstaates gefahrenen Meilen.

Or they can choose **not** *to have a device at all, opting instead to pay a flat fee based on the average number of miles driven by all state residents.*

Constraints on Argument Features

18 of the explicit CHOSEN ALTERNATIVE in the PDTB (15.3%) and 24 of the implicits (14.0%) have an **downward-entailing** expression in **Arg1**

(17) **In India, he rejects the identification of Indianness with Hinduism, ...** Instead **he champions Mr Tagore's view . . . .** [The Economist, 18 June 2005]

John **rejected** dogs. ⇓ John **rejected** beagles.

(18) **The current system is too bureaucratic . . . .** Instead, **research councils should "pay the full costs of the projects they fund . . . "**. [Research Fortnight, 28 April 2004]

John was **too ill** to own a dog. ⇓ John was **too ill** to own a beagle.

Constraints on Argument Features

Another 14 of the explicits (11.9%) and 9 of the implicits (5.3%) have a **modal** marker of either obligation, possibility, desire, etc. in **Arg1**.

(19) **Charles Kennedy's advisors should have told him the truth**. Instead, **they covered up for him to an unacceptable extent and for far too long**. [The Economist, 14 January 2006]

(20) **Anne Compoccia wanted to be a nun**. Instead, **she found herself in prison for embezzling city funds**.

[http://www.nytimes.com/2002/12/22/nyregion/22DECA.html?todaysheadlines]

Constraints on Argument Features

Again, while I don't have counts, modals are found in **Arg1** of similar relations in German.

(21) Auch **sollten** wir sie nicht zu kritisch sehen. <u>Stattdessen</u> sollte unser Ziel darin bestehen, uns auch abseits der ständigen Aktivitäts- und Technologieberieselung wohlzufühlen. [wmt13]

Nor **should** *we be too critical of it. <u>Rather</u>, our goal should be to feel comfortable away from the constant chatter of activity and technology.*

(22) Dieser sympathische und intelligente, aber vom Leben gestrafte junge Mann **könnte** als Touristenführer, Kellner oder am Empfang eines Hotels arbeiten, aber <u>stattdessen</u> verrichtet er die Arbeit eines Tragesels.

Punished for life, this pleasant, intelligent young man **could** *be a tour guide, waiter or hotel receptionist, but instead he does the work of a mule.*

Constraints on Argument Features

Just these three features (negation, a DE expression, or an event modal) account for 78 (66.1%) of the explict tokens of CHOSEN ALTERNATIVE and 149 (87.1%) of the implicit tokens in the PDTB.

Further analysis should turn up other features.

These features can be used to prefer translations that contain one of them to translations which don't.

Constraints on Argument Features

SRC: Prof Vallortigara said he did**n't** think that the dogs were intentionally communicating with each other through these movements. **Instead**, he believes that the dogs have learned from experience what moves they should and shouldn't feel worried about.

REF: Nach Ansicht von Professor Vallortigara kommunizieren die Hunde **nicht** absichtlich miteinander durch diese Bewegungen. Er ist **vielmehr** berzeugt, dass die Hunde aus Erfahrung gelernt htten, bei welchen Bewegungen sie sich Sorgen machen sollten und wann nicht.

Constraints on Argument Features

MT: Prof Vallortigara sagte, er glaube, dass die Hunde absichtlich miteinander zu kommunizieren, durch diese Bewegungen. **Stattdessen** glaubt er, dass sie Hunde aus Erfahrung gelernt haben, was sie bewegt und beunruhigt fühlen sollte.

GLOSS: Prof Vallortigara said that he believed that the dogs intentionally to communicate with each other by these movements. **Instead**, he believes that dogs have learned from experience, what moves them and should feel worried.

Final Problem: Language-specific differences in usage

Explicit discourse connectives are less frequent in Chinese than in English [Zhou & Xue, forthcoming]:

- The Chinese Discourse TreeBank contains 4x as many implicit discourse relations as explicitly marked relations.
- The Penn Discourse TreeBank contains about equal numbers.

Explicit disourse connectives are used less frequently in English than in German [Becher, 2011]:

- In translating into German a corpus of English corporate reports (letters to shareholders and mission statements, ∼21K words), translators **added** 114 connectives and **dropped** 32.
- In translating into English a corpus of similar German corporate reports (∼21K words), translators **added** only 48 connectives but **dropped** 51.
- Counts were not given for connectives in the original corpora.

We need ways of handling these omissions and additions in building translation models.

Conclusion

- From a linguistic perspective, I think there are several aspects of discourse that can be easily addressed in SMT.
- Other aspects, such as languages differing in what they need to make explicit, raise problems known from sentence-level semantics.
- The technical challenges of enlarging the locality over which translation hypotheses can be ranked and decisions made, are ones that several people seem willing — even eager — to address.

References

[Becher, 2011]
Viktor Becher. When and why do Translators add connectives? A corpus-based study. *Target* 23:1, pp. 26–47, 2011.

[Foster, Isabelle & Kuhn, 2010]
George Foster, Pierre Isabelle, and Roland Kuhn. Translating structured documents. *Proceedings of AMTA*, Denver CO, 2010.

[Guillou, 2012]
Liane Guillou. Improving pronoun translation for statistical machine translation. *Proceedings, European Chapter of the Association for Computational Linguistics (EACL)*, 2012.

[Guillou et al, 2014]
Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann and Bonnie Webber. ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT. Proc. 9th International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, pp. 3191–3198, 2014.

[Hajlaoui & Popescu-Belis, 2013]
Najeh Hajlaoui & Andrei Popescu-Belis. Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric. *Proc. CICLing 2013* (14th International Conference on Computational Linguistics and Intelligent Text Processing), LNCS 7817, 2013, pp.236-247.

[Hardmeier & Federico, 2010]
Christian Hardmeier and Marcello Federico. Modelling pronominal anaphora in statistical machine translation. *Proc. 7th Int'l Workshop on Spoken Language Translation*, pages 283–290, 2010.

[Le Nagard & Koehn, 2010]
Ronan Le Nagard and Philipp Koehn. Aiding pronoun translation with co-reference resolution. *Proc. 5th Joint Workshop on Statistical Machine Translation and Metrics (MATR)*, 2010.

[Li, Carpuat & Nenkova, 2014]
Junyi Jessy Li, Marine Carpuat and Ani Nenkova. Assessing the Discourse Factors that Influence the Quality of Machine Translation. Proc. 52nd Annual Meeting of the Association for Computational Linguistics, pp. 283–288, 2014.

[Louis & Webber, 2014]
Annie Louis and Bonnie Webber. Structured and Unstructured Cache Models for SMT Domain Adaptation. *Proc. European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, 2014, pp. 55–163.

[Meyer & Popescu-Belis, 2012]
Thomas Meyer and Andrei Popescu-Belis. Using Sense-labeled Discourse Connectives for Statistical Machine Translation, *Proc. Workshop on Hybrid Approaches to Machine Translation (HyTra)*, 2012, pp.129–138.

[Novák, 2011]
Michal Novák. Utilization of Anaphora in Machine Translation. *WDS 2011*, Week of Doctoral Students, June 2011. (http://ufal.mff.cuni.cz/michal-novak)

Pitler & Nenkova [2009]
Emily Pitler & Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. *ACL-IJCNLP '09*, Singapore, 2009, pp.13–16.

[Prasad et al, 2008]
Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Aravind Joshi and Bonnie Webber. The Penn Discourse Treebank 2.0. *Proc. 6th International Conference on Language Resources and Evaluation*, 2008.

[Prasad et al, 2014]
Rashmi Prasad, Bonnie Webber and Aravind Joshi. Reflections on the Penn Discourse TreeBank, Comparable Corpora and Complementary Annotation. Computational Linguistics, doi:10.1162/COLI_a_00204.

[Rysova 2012]
Magdaléna Rysová Alternative lexicalizations of discourse connectives in Czech. *Proc. 8th Int'l Conf. Language Resources and Evaluation*, 2012, pp. 2800–2807.

[Webber, 2013]
Bonnie Webber. What excludes an alternative in coherence relations? *Proc. 10th International Conference on Computational Semantics (IWCS 2013)*, pp. 276–287, Potsdam, Germany.

[Zhou & Xue, forthcoming]
Yuping Zhou and Nianwen Xue. The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. *Journal of Language Resources and Evaluation*, to appear.