Discourse and SMT

Bonnie Webber
School of Informatics
University of Edinburgh

September 13, 2012

1. Introduction
2. Aspects of discourse relevant to SMT
3. How discourse can contribute to SMT
4. Conclusion

Overview

- Discourse involves information conveyed by **segments** larger than a single clause.
- **Sentences** are segments with $\geq$1 clauses; Sequences of sentences always involve $>$1 clause.
- People must be able to recognize and extract information about these segments without too much added effort.
- All languages provide **devices** that allow people to do this, using the **context** they construct from the discourse.
- The devices vary from language to language.
- SMT can be sensitive to discourse, even when the unit of translation is a single sentence.

Aspects of discourse relevant to MT: Segments

Discourse segments larger than a clause may be defined in terms of

- the particular **topic** the segment addresses or the particular **function** it fulfills;
- the relation(s) between their constituent clauses/sentences.

Aspects of discourse relevant to MT: Segments

⇒ Segments on the same topic or with the same function can resemble each other in terms of the words and/or syntax they use.

⇒ Relations between the clauses/sentences in a segment may be signalled through

- how they combine
- their individual structure (e.g., parallel structure)

Aspects of discourse relevant to MT: Clause-combining

Clauses may be combined into segments that relate to each other via **coordinating conjunctions** or **adjacency**:

(1) I don't kill flies **but** I like to mess with their minds. I hold them above globes. They freak out and yell, 'Whoa, I'm way too high!'. [Bruce Baum]

or **subordinating conjunctions**, or **discourse adverbials**:

(2) 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]

Additional meaning is conveyed through how clauses combine.

Aspects of discourse relevant to MT: Clause-combining

Clauses can combine in different ways, across ≥1 sentences, while conveying the same meaning (i.e., **paraphrase**):

(3) a. The market for export financing was liberalized in the mid-1980s, forcing the bank to face competition.

b. The market for export financing was liberalized in the mid-1980s, which forced the bank to face competition.

c. When the market for export financingwas liberalized in the mid-1980s, it forced the bank to face competition.

d. The liberalization of the market for export financing forced the bank to face competition.

e. The market for export financing was liberalized in the mid-1980s. This forced the bank to face competition.

Aspects of discourse relevant to MT: Contextual devices

Languages have devices that exploit the **context** of the previous text to allow information to be conveyed with minimal effort.

Minimal effort might involve **an expression of coreference**:

(4) The police are not here to create disorder. **They** are here to preserve **it**. [Attributed to Yogi Berra]

(5) What if everything is an illusion and nothing exists? In **that case**, I definitely overpaid for my carpet. [Woody Allen]

or **sentence fragments**:

(6) Pope John XXIII was asked "How many people work in the Vatican?". He is said to have replied, "About **half**".

Aspects of discourse relevant to MT: Contextual devices

As with clause combining, different contextual devices can express the same meaning.

(7) Pope John XXIII was asked "How many people work in the Vatican?". **The Pope** is said to have replied, "About half".

(8) When asked "How many people work in the Vatican?", Pope John XXIII is said to have replied, "About half".

Outline

- A bit more detail on aspects of discourse relevant to SMT:
  - Topic structure and segmentation
  - Functional structure and segmentation
  - "Clause combining" and discourse relations
  - Contextual devices
- What's be done to use them in SMT

Topic Structure and Segmentation

Expository text can be viewed as a sequence of **topically coherent** segments. Their 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

Topic Structure and Segmentation

Being able to recognize topic structure was originally seen as benefitting **information retrieval** [hea97]

Recent interest comes from its use in **segmenting lectures or other speech events**, making them more amenable to search [gal03,mal06].

Topic Structure and Segmentation

Techniques for topic segmentation assume:

- the topic of each segment differs from those of its adjacent sisters;
- adjacent spans that share a topic belong to the same segment;
- topic predicts lexical choice, either of all words of a segment or just its content words (ie, excluding "stop-words").

Topic Structure and Segmentation

Techniques for topic segmentation make use of either:

- **semantic-relatedness**, where words within a segment are taken to relate to each other more than to words outside the segment [hea97,choi01,bes06,gal03,mal06]
- **topic models**, where each segment is taken to be produced by a distinct, compact lexical distribution [chen09,eis08,purv06]

An excellent overview and survey of this work can be found in [purv11].

Functional Structure and Segmentation

Texts within a given genre – eg,

- news reports
- scientific papers
- letters to the editor of a newspaper, magazine, journal, etc.
- . . .

generally share a similar structure, that is independent of topic (eg, sports, politics, disasters; or molecular genetics, radio astronomy, SMT), instead reflecting the **function** played by their parts.

Example: News Reports

Functional structure of news reports is an **inverted pyramid**:

- **Headline** gets the reader's attention
- **Lead paragraph** (sometimes spelled *lede*) conveys **who** is involved, **what** happened, **when** it happened, **where** it happened, **why** it happened, and (optionally) **how** it happened
- **Body** provides more detail about who, what, when, . . .
- **Tail** contains less important information

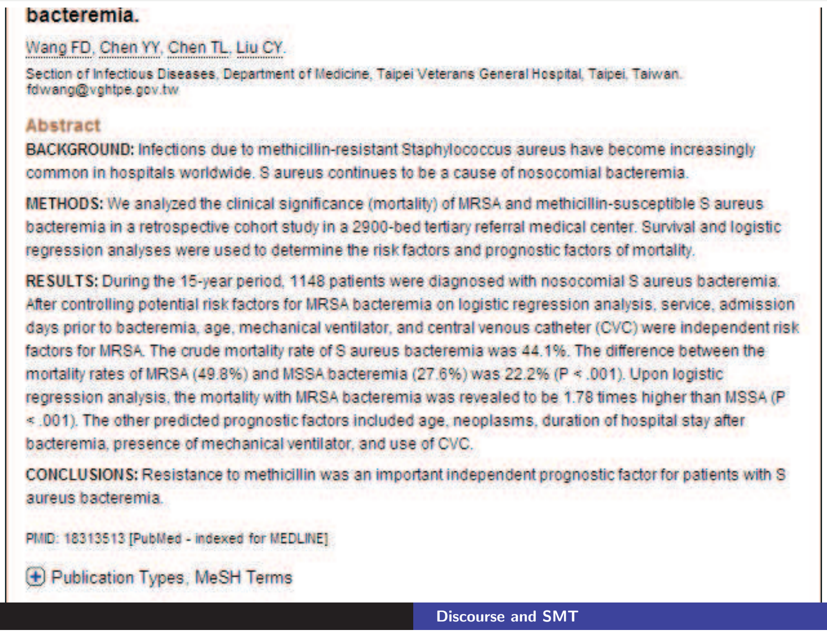
This structure is why the first (ie, lead) paragraph is usually the best *extractive summary* of a news report.

Example: Scientific articles/abstracts

The functional structure of **scientific articles** comprises:

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

Abstracts with a similar structure are called **structured abstracts**.



bacteremia.

Wang FD, Chen YY, Chen TL, Liu CY.

Section of Infectious Diseases, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan.
fdwang@vghtpe.gov.tw

Abstract

BACKGROUND: Infections due to methicillin-resistant Staphylococcus aureus have become increasingly common in hospitals worldwide. S aureus continues to be a cause of nosocomial bacteremia.

METHODS: We analyzed the clinical significance (mortality) of MRSA and methicillin-susceptible S aureus bacteremia in a retrospective cohort study in a 2900-bed tertiary referral medical center. Survival and logistic regression analyses were used to determine the risk factors and prognostic factors of mortality.

RESULTS: During the 15-year period, 1148 patients were diagnosed with nosocomial S aureus bacteremia. After controlling potential risk factors for MRSA bacteremia on logistic regression analysis, service, admission days prior to bacteremia, age, mechanical ventilator, and central venous catheter (CVC) were independent risk factors for MRSA. The crude mortality rate of S aureus bacteremia was 44.1%. The difference between the mortality rates of MRSA (49.8%) and MSSA bacteremia (27.6%) was 22.2% (P < .001). Upon logistic regression analysis, the mortality with MRSA bacteremia was revealed to be 1.78 times higher than MSSA (P < .001). The other predicted prognostic factors included age, neoplasms, duration of hospital stay after bacteremia, presence of mechanical ventilator, and use of CVC.

CONCLUSIONS: Resistance to methicillin was an important independent prognostic factor for patients with S aureus bacteremia.

PMID: 18313513 [PubMed - indexed for MEDLINE]

Functional Structure and Segmentation

Automatic annotation of functional structure is seen as benefitting:

- **Information extraction**: Certain types of information are likely to be found in certain sections [Moe99,Moe00]
- **Extractive summarization**: More "important" sentences are more likely to be found in certain sections.
- **Sentiment analysis**: Words that have an objective sense in one section may have a subjective sense in another [tab09]
- **Citation analysis**: A citation may serve different functions in different sections [teu10]

Functional structure and segmentation

Techniques for functional segmentation assume:

- Function predicts more than lexical choice:
  - indicative phrases such as "results show" (→ *Results*)
  - indicative stop-words such as "then" (→ *Methods*).
- Functional segments usually appear in a specific order, so either sentence position is a feature in the models or sequential models are used.

Functional structure and segmentation

Many biomedical journals made **structured abstracts** mandatory in late 1990 / early 2000.

Before that, structure was rarely indicated explicitly.

Assuming that the writing of abstracts didn't change — just the addition of section labels, this led to much of the early work on functional segmentation being on biomedical text, where abstracts with labelled sections were taken as **training data** for segmenting unlabelled abstracts [chu09,guo10,hir08,lia10,lin06,mckn03,ruch07],

More recent work on functional segmentation has involved **meeting transcripts**, both for indexing and summarization [Nie09; Nie12]

Discourse relations and low-level segmentation

Discourse relations produce a low-level (possibly overlapping) discourse segmentation. This requires identifying

1. the evidence for a relation between discourse elements (clauses and/or sentences);
2. the discourse elements being related;
3. the type(s) of sense relation that hold(s) between them.

Discourse relations and low-level segmentation

The general problems are:

1. Given a language, what are its standard ways of combining clauses? (Languages like Danish and Arabic tend to favor coordination, while Italian favors subordination.)
2. Since devices used to combine clauses or other discourse elements may be ambiguous, when does a token in text serve that role?
3. Given a token that does relate discourse elements, which ones does it relate (ie, which serve as its arguments)?
4. Given such a token and its arguments, what sense relation(s) hold between the arguments?

Discourse relations and low-level segmentation

When does an individual token serve to combine clauses and signal a discourse relation, since they are often *syntactically ambiguous* [pit09b]:

(9) Asbestos is harmful **once** it enters the lungs. *(subordinating conjunction)*

(10) Asbestos was **once** used in cigarette filters. *(adverb)*

Surface cues allow discourse and non-discourse use to be distinguished with at least 94% accuracy [pit09b].

Discourse relations and low-level segmentation

Given a token that serves to combine clauses and relate discourse elements, which does it combine as its arguments?

So far, no language has shown discourse connectives that relate more or less than two arguments:

- **Arg2** – argument syntactially bound to the connective
- **Arg1** – the other argument

Discourse relations and low-level segmentation

With **Arg2**, the main question is whether any *attribution* it may contain is included in the argument.

(11) **We pretty much have a policy of not commenting on rumors**, and I think **that falls in that category**. [wsj_2314]

(12) **Advocates said** **the 90-cent-an-hour rise, to $4.25 an hour by April 1991, is too small for the working poor**, while **opponents argued** **that the increase will still hurt small business and cost many thousands of jobs**. [wsj_0098]

Discourse relations and low-level segmentation

With **Arg1**, identification is harder because it need not be adjacent to **Arg2**:

(13) On a level site you can provide a cross pitch to the entire slab by **raising one side of the form** (step 5, p. 153), but for a 20-foot-wide drive this results in an awkward 5-inch (20 x 1/4 inch) slant across the drive's width. Instead, **make the drive higher at the center**.

(14) **Big buyers like Procter & Gamble say there are other spots on the globe and in India, where the seed could be grown**. "It's not a crop that can't be doubled or tripled," says Mr. Krishnamurthy. But **no one has made a serious effort to transplant the crop**. [wsj_0515]

Discourse relations and low-level segmentation

Methods to identify the args to a discourse relation include:

- a **discriminative log-linear ranking model** on syntactic, dependency and lexical features, to separately identify connectives and their arguments [wp07], plus a **log-linear re-ranking model** to select the best **pair** of arguments, to capture dependencies between them.

Type of connective	Ranking Accuracy	Re-ranking Accuracy
Coordinating conjunctions	75.5%	78.3%
Subordinating conjunctions	87.2%	86.8%
Discourse adverbials	42.2%	49%

⇒ Dependencies between the args of **coord conjunctions** and **discourse adverbials**, but not between args of **subord conjunctions**.

Discourse relations and low-level segmentation

Other methods include:

- **connective specific models** [elw08], which improves recognition of args to **discourse adverbials** (from 49.0% to 67.5%), while degrading performance for **subord conjunctions** and doing nothing for **coord conjunctions**
- **location specific methods** [pra10], where Arg1 of an **inter-sentential connective** is in the same paragraph as Arg2 4301/4373 = 98% of the time, and the average WSJ paragraph has only 3 sentences.

Identifying the sense of a discourse relation

Given a set of sense labels, one wants to choose the one or more that hold in a given instance.

Some **explicit** discourse connectives are unambiguous with respect to sense:

Conn	sense	Conn	sense
accordingly	RESULT (5/5)	in addition	CONJUNCTION (165/165)
additionally	CONJUNCTION (7/7)	moreover	CONJUNCTION (100/101)
afterward	PRECEDENCE (11/11)	so	RESULT (262/263)
as a result	RESULT (78/78)	thus	RESULT (112/112)
consequently	RESULT (10/10)	till	PRECEDENCE (3/3)
for instance	INSTANTIATION (98/98)	unless	DISJUNCTIVE (94/95)

Classifying marked sense relations

But several common connectives can express ≥1 sense:

- **since**: REASON (94), SUCCESSION (78)
- **as**: SYNCHRONY (387), REASON (166)
- **and**: RESULT (38), CONJUNCTION (2543), both of these simultaneously (138)

[pit09b] trained a simple Naive Bayes classifier to 94.15% accuracy in disambiguating between whether an explicit connective expressed one of four high-level senses (CONTINGENCY, TEMPORAL, COMPARISON, EXPANSION) based on lexical and syntactic features.

Classifying unmarked relations

Where there is no explicit discourse connective, evidence for the relation may be derivable from other features.

(15) [A car had broken down on an unmanned level crossing and was hit by a high speed train.]
[The train derailed.]
→ **Result**

(16) [The damage to the train was substantial,]
[fortunately nobody was injured]
→ **Contrast**

Classifying unmarked relations: Marcu and Echihabi (2002)

But considerable training data is needed, since the features are sparse.

Data:

- 4 sense relations from RST [mt87]: contrast, condition, cause-explanation-evidence, elaboration;
- 2 non-relations: no-rel-same-text, no-rel-different-text;
- 900,000 to 4 million automatically labelled examples per relation, derived from clauses connected by unambiguous subord or coord conjunctions.

Model:

- Naive Bayes
- Word co-occurence features as predictors of the relation indicated by the clauses are conjoined.

Classifying unmarked relations: Marcu and Echihabi (2002)

Results:

- test on automatically labelled data: 49.7% accuracy for 6-way classifier
- test on manually labelled examples from RST TreeBank [car03] without removing discourse connectives from training data and using binary classifiers: 63% to 87% accuracy
- test on manually labelled, unmarked examples using binary classifiers (contrast vs. elaboration, and cause-explanation-evidence vs. elaboration): 69.5% recall for contrast, 44.7% recall for cause-explanation-evidence

Subsequent work has shown that it's worth making use of more features, and that marked relations differ from ones "born unmarked".

Contextual devices

Police are not meant to create disorder. **They** *are meant to preserve* **it**.

- "Police", "disorder", "they", and "it" are **referring expressions**.
- Expressions like "Police" and "disorder" lead to entities (their **referents**) entering into the **context**, which can then be used to interpret the subsequent text.
- The personal pronouns "they" and "it" are **anaphoric expressions**, which rely on **context** for their interpretation.
- Personal pronouns rely on context by **coreferring** to a **referent** already in the model.

Contextual devices

Other contextual devices rely on context in other ways than coreference:

- **fragments**

  (17) Pope John XXIII was asked "How many people work in the Vatican?". He is said to have replied, "About **half**".
- **comparative anaphors** like "other".

  (18) **Other contextual devices** include comparative anaphors.
- **verb phrase ellipsis** (VPE)

  (19) Fred doesn't like football, but Mary does.

  (20) You can go on Monday, but Tuesday you can't.

Where discourse features can contribute to SMT

Discourse suggests that we can take advantage of:

- similar words and/or syntax being found in segments on the same topic or with the same function;
- finding different ways to combine clauses in the source text, that more closely resemble the target or are easier to translate;
- disambiguating ambiguous discourse connectives in a source text, to better map them into the target.
- recognizing the sense of implicit discourse connectives in a source text, to explicitate them in the target.
- resolving contextual devices (pronouns, VPEs) in a source text, in order to realize them correctly in the target.

Discourse Segmentation and SMT

○ Methods for topic and functional segmentation rely on topic predicting lexical choice (and syntactic choice, in the latter case).

⇒ Foster, Isabelle & Kuhn (2010) explore whether, by

- characterizing segments, and
- producing a different Language Model for each segment type

SMT can be improved through assuming that the language used in segments of a given type (but from different documents) is a more accurate Language Model than a model of more general language.

Discourse Segmentation and SMT

○ Foster et al. use a Hansard corpus of transcripts of Canadian parliamentary proceedings.

○ Each "document" comprises a sequence of contributions from several speakers, each contribution associated with a particular parliamentary activity and daily parliamentary routine.

○ As such, each segment (and each of its sentences) can be characterized by:

- **session**: a year between 2001 and 2009
- **source language**: English or French
- **speaker**: 586 names, with a Zipfian distribution over their volume of contributions
- **title**: 45 parliamentary activities, with *Debate* most common
- **section**: 4 general types of daily routines

Discourse Segmentation and SMT

○ Foster et al. develop specific models (in English and in French) for each feature value, with feature-specific models used to produce the best translation hypothesis for each source sentence.

○ In terms of BLEU scores, this produces a modest, but statistically significant improvement, in both translation directions.

⇒ Can automated segmentation (by topic, function, . . .) of some corpus in need of translation produce similar or greater benefits?

Discourse relations and SMT

- Disambiguating markers of discourse relations for SMT
- Identifying clause-combining patterns for sentence-alignment in SMT

Discourse relations and SMT

1. Discourse connectives may cover different sense spaces in different languages.

- *Since* in English can express either an explanation (like *because*) or a temporal relation (like *after*).
- *Puisque* in French expresses only the former sense, while *depuis* expresses only the latter.

⇒ Work by Meyer and colleagues at Idiap suggests that recognizing and annotating **relational structures** in the source can allow appropriate discourse connectives to be selected in the target.

Discourse relations and SMT

2. Translators often make discourse connectives **explicit** in their target translation that were implicit in the source [KO11]

Connective	Orig Frequency	Trans Frequency
therefore	0.153%	0.287%
nevertheless	0.019%	0.045%
thus	0.015%	0.041%
moreover	0.008%	0.035%

⇒ This can produce source-target mis-alignments that produce bad entries in the translation model.

Discourse relations and SMT

Explicitating **implicit connectives** in the **source text** should improve alignment and thus SMT.

Hypothesis: Although recognizing coherence relations that hold between otherwise unmarked sentence pairs is hard in general, it might be simpler for those connectives that get explicitated.
E.g. Implicit THEREFORE and THUS:

(21) **Its valuation methodologies**, she said, "**are recognized as some of the best on the Street**. Implicit = THEREFORE **Not a lot was needed to be done**." [wsj_0304]

(22) "**In Asia, as in Europe, a new order is taking shape**," Mr. Baker said. Implicit = THUS "**The U.S., with its regional friends, must play a crucial role in designing its architecture**." [wsj_0043]

Discourse relations and SMT

3. Patterns of clause-combining could prove useful for **splitting** sentences that **do not** participate in 1:1 alignments, or to produce a sequence of shorter sentences.

5-10% of sentences in bi-texts are **discarded** because they do not participate in 1:1 alignments.

(23) Sometimes it is worthy of satire and merits discussion, but I digress.

(24) Manchmal ist das schon kabarettreif und verdient eine Diskussion.
Das ist aber nicht mein Punkt.

(25) This is important, but so is enforcement and there are, of course, a number of reasons why we need to pay particular attention to this.

(26) Das halte ich ebenso wie die Umsetzung für wichtig.
Natürlich gibt es gute Gründe, weshalb wir diesem Problem besondere Aufmerksamkeit widmen müssen.

Discourse context and SMT

- Pronoun anaphora
- Verb Phrase Ellipsis

Pronoun Anaphora and SMT

Anaphors — in particular, pronouns and 0-anaphors — are constrained by their **antecedents** in all languages, but in different ways.

- English: Pronoun gender reflects the **referent** of the antecedent.
- French, German, Czech: Pronoun gender reflects the **form** of the antecedent.

(27) a. Here's a book. I wonder if **it** is new. (inanimate, neuter referent)

b. Voici un livre. Je me demande si **il** est nouveau. (masculine form)

Discourse context and SMT: Pronoun anaphora

Phrase-based and syntax-based SMT just consider the local context - cf. Google Translate (as of 15 July 2012)

(28) Mary has **a book**. I wonder if **it** is new.
GT: **Marie** a **un livre**. Je me demande si **elle** est nouvelle.

(29) John had **an orange**. I wondered if **it** was new.
GT: **John** avait **une orange**. Je me demandais si **il** était neuf.

⇒ **elle** can only refer to feminine: **un livre** is masculine.
⇒ **il** can only refer to masculine: **une orange** is feminine.

Pronoun Anaphora and SMT

Hypothesis: Co-reference resolution on the source language text may enable appropriate forms to be chosen in the target language.

Preliminary work has been done on this by

- Le Nagard & Koehn (2010): English–French [NK10]
- Hardmeier & Federico (2010): English–German [HF10]

both using effectively the same procedure.

Pronoun Anaphora and SMT

What the procedure does is:

1. Identify source language pronouns that **co-refer**. (Other pronouns are purely syntactic.)

   It *is raining.*
   It *is possible that we will arrive late.*

2. For each co-referring pronoun, use anaphor resolution to identify its antecedent NP;
3. Identify the syntactic head of that NP;
4. Locate the alignment of the head in the target text.
5. Identify relevant features of the aligned element;
6. Annotate the source text pronoun with these features.

Pronoun Anaphora and SMT

The resulting **annotated** source language text is used to train a Translation Model.

In order to use this enriched TM in translation,

1. each co-referring source text pronoun is first resolved, prior to translation.
2. During the translation process, the translation of each pronoun's antecedent must be identified.
3. Appropriate features must be extracted from the translation and those features annotated onto the source text pronoun.
4. Then the sentence, with its pronouns annotated, can then be translated.

Pronoun Anaphora and SMT

Earlier results showed a small but disappointing improvement:

○ LeNagard & Koehn: Manual evaluation of a subset.

- 40/59 pronouns annotated (68%), with 33/59 annotated correctly (56%)
- 27/33 of those correctly translated (82%)
- 41/59 pronouns correctly translated in baseline (69%)

○ Hardmeier & Federico: Automated approximate recall & precision (ie, presence of pronouns in both source and translated text)

- Baseline F-score: 31.7% on 2008 WMT test set, 40.7% on 2009 test set
- Pronoun model F-score: 32.6% on 2008 test set, 41.4% on 2009

Pronoun Anaphora and SMT

Guillou [Gui12] substituted a set of gold standard English–Czech corpora to see why the procedure led to so small an improvement.

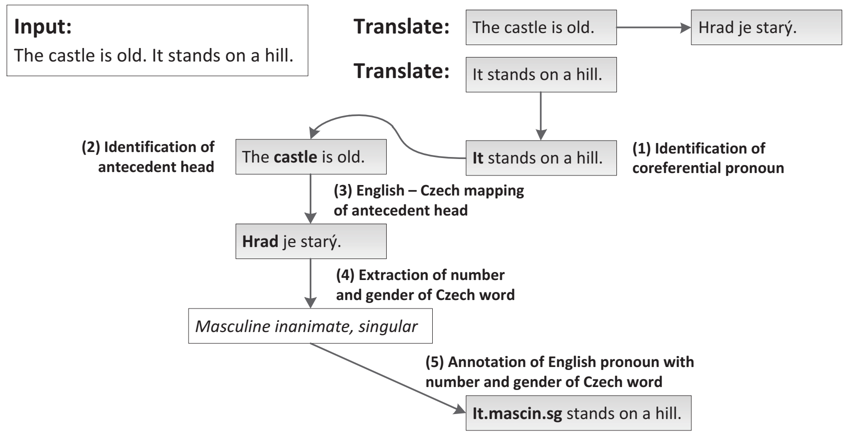
3^{rd}-person Czech pronouns: masculine (animate and inanimate), feminine, neuter.

(30) The **dog** has a ball. I can see **it** playing outside.
dog = pes (masculine, animate)
it = ho

(31) The **cow** is in the field. I can see **it** grazing.
cow = kráva (feminine)
it = ji

(32) The **car** is in the garage. I will take **it** to work.
car = auto (neuter)
it = ho

Annotation Projection


Input:
The castle is old. It stands on a hill.
Translate:
The castle is old.
Hrad je starý.
Translate:
It stands on a hill.
(2) Identification of antecedent head
The **castle** is old.
It stands on a hill.
(1) Identification of coreferential pronoun
(3) English – Czech mapping of antecedent head
Hrad je starý.
(4) Extraction of number and gender of Czech word
Masculine inanimate, singular
(5) Annotation of English pronoun with number and gender of Czech word
It.mascin.sg stands on a hill.

Potential Sources of Error

The process assumes that errors arise when:

- Deciding whether or not a third person pronoun corefers;
- Identifying the pronoun antecedent;
- Identifying the head of the antecedent;
- Aligning the source and target texts at the phrase and word levels.

Results

- Improvement over the Baseline, but only very small
- Improvement not statistically significant due to small datasets
- Did not meet expectations - investigation required

Possible Sources of Error

Other sources of error:

- Mis-Identification of the English antecedent head noun
- Mis-Identification of the Czech translation of the antecedent head
- Errors in the PCEDT 2.0 alignment file (affecting training only)

Next steps

- **Is source-side annotation enough?**
  - Do we keep, remove, or combine it with something else?
- **Automated evaluation metrics remain the holy grail**
- **Should the problem be viewed as translating pronouns or as expressing coreference?**
- **Paraphrase techniques for generating synthetic reference translations**

VPE translation in English-French SMT [Leirvik, 2012]

What is VPE?

Verb phrase ellipsis (VPE) "occurs when an auxiliary or modal verb abbreviates an entire verb phrase found elsewhere in the context." [BS11]

- *She doesn't like the film, but he does ~~like the film~~.*
- *You can go on Monday, but you can't ~~go~~ on Tuesday.*

VPE in SMT

- VPE is a common syntactic construction in English that is rare in other languages.
- To translate VPE in English source text,
  - tokens must first be detected,
  - then something must be generated in its stead: its **antecedent** or some **reduced form** or some **idiomatic construction**
- Detecting VPE requires syntactic information
  - not available in the standard phrase-based SMT approach
- Successful handling of VPE may also require identifying long-range dependencies if they have to be resolved to be translated.

How well does Google do?

VPE with a subject pronoun:

He doesn't want to speak, so she will.
Il ne veut pas parler, alors elle le fera.

VPE with a full NP subject:

He doesn't want to speak, but the woman in the hat does.
Il ne veut pas parler, mais la femme dans le chapeau fait.

VPE detection

- Hardt [Har93]
  - Penn Treebank: Use syntactic patterns: 31% precision
  - Brown Corpus: Use string-matching and POS tags: 45% precision
- Nielsen [Nie04]
  - Manually identified a training set of VPE instances
  - Trained a MaxEnt classifier on surrounding words/POS tags + other features (56–79% precision depending on features used)
- Bos & Spenader [BS11]
  - Manually annotated Penn Treebank for VPE instances and their antecedents
  - A gold standard for all future work!

French translation strategies

- On the target side, 92 (20%) translations of VPE into French explicitly use the antecedent VP.
- But $\geq$50% use a common reduced form:

Strategy	Example	Frequency
SUBJ+BE IT	Certaines sont bonnes et *certaines ne le sont pas.*	58
SUBJ+DO IT	Je pourrais citer des pays, *je ne le ferai pas.*	47
IT+BE THE CASE	Vous auriez pu r'egler tout ceci mais *cela n'a pas été le cas.*	39
SUBJ + NOT	Nous avons le temps, *Saddam pas.*	13
⋮	⋮	⋮

- The rest use a reduced form that is specific to that translation.

Systematic evaluation of VPE translation

- How do we know if the VPE has been translated correctly?
  - BLEU is no help!
- Subjective assessment is costly.
- Idea: Use the corpus itself to identify other possible correct translations [OGVW06]:
  1. Group VPE instances by English class and subject pronoun, replacing any full NP subject with an appropriate pronoun;
  2. Create a list of corresponding French translation strategies;
  3. This expands the set of reference translations for any new instance from that English class.

Evaluation example

- Test sentence: Some Member States operate a card system, *others do not*.
- Collect all VPE instances containing *[they] do not*

Some countries ban organisations, *others*$_{(they)}$ *do not*.	→	Certains pays inderdisent ces organisations, alors que *d'autres non*.	SUBJ+NO
Animals have rights, *children*$_{(they)}$ *do not*.	→	Les animaux ont des droits, *les enfants pas*.	SUBJ+NOT
Those large ones employ staff and *the small ones*$_{(they)}$ do *not*.	→	Celles-ci emploient des travailleurs, *ce qui n'est pas le cas des petits*.	WHICH+BE THE CASE+SUBJ
	⋮		

Evaluation example

- Possible translations:
  - Certains états members pratiquent un système de cartes, *d'autres non*.
  - Certains états members pratiquent un système de cartes, *d'autres pas*.
  - Certains états members pratiquent un système de cartes, *ce qui n'est pas le cas d'autres*.

  . . .

Baseline results

- Trained 377 automatically detected VPE instances (of which 321 correct)
- Tested on 166 instances (of which 136 correct)
- Of the true VPE instances:
  - 10 match the reference VPE translation
  - 25 use a correct alternative
  - 101 are wrong

Related work: Null elements in SMT

Chung and Gildea (2010) looked at improving translation from pro-drop languages (Chinese and Korean) into English.

- Added a null element to source language sentences at each empty pronoun position.
- Improved BLEU score by 1 point

VPE placeholder insertion results

Using the approach of Chung and Gildea (2010), we get

- 46 instances are translated correctly
- 89 are wrong
  - Including 15 which were translated correctly by the baseline system
- Overall, 12% improvement over the baseline system
- Most of the correctly translated instances involve a modal verb in the English VPE phrase
- The most common English VPE classes (those involving BE and DO are still being translated incorrectly.
  - In fact, the "improved" system does worse on these classes than the baseline does!

Error analysis I

- The most common incorrect translation is SUBJ+AUX+"the case"
  - I know that one of the projects is included at the other is not [VPE]. → Je sais que l'un des projets est inclus et *les autres n'est pas le cas.**
- The most common correct translation is SUBJ+AUX+"it"
  - People in close proximity can come, other people can not [VPE]. → Les gens à proximité peut venir, *d'autres personnes ne le pouvons pas*.
- SUBJ+AUX+"IT" and "it"+"be the case" (or similar) occur in nearly all English VPE classes – other strategies are more limited

Error analysis II

A placeholder only a good idea in SMT when

- The target sentence contains everything in the source sentence, plus something else.
- You can predict correctly where the placeholder should go (ie, where that something else is!).
- You can force what the placeholder is translated with.

Conclusion

- Discourse has several properties that are relevant to the quality of SMT.
- Even if SMT operates at the level of the sentence, it's possible to reflect properties of discourse.

References

[bes06] Yves Bestgen (2006). Improving text segmentation using Latent Semantic Analysis: A reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, 32(1):5–12.

[BS11] Bos, J. and Spenader, J. (2011). An annotated corpus for the analysis of VP ellipsis. *Language resources and evaluation, 45*(4), 463–494.

[car03] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt & R. Smith, editor, *Current Directions in Discourse and Dialogue*. Kluwer, New York.

[Chen09] Harr Chen, S. R. K. Branavan, Regina Barzilay, and David Karger (2009). Global models of document structure using latent permutations. *Proc. NAACL'09*, 371–379.

[choi01] Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore (2001). Latent Semantic Analysis for text segmentation. *Proc. EMNLP'01*, 109–117.

[chu09] Grace Chung (2009). Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*, 10(9).

[CG10] Chung, T. and Gildea, D. (2010). Effects of empty categories on machine translation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 636–645).

[eis08] Jacob Eisenstein and Regina Barzilay (2008). Bayesian unsupervised topic segmentation. *Proc. EMNLP*, pages 334–343.

[elw08] Robert Elwell and Jason Baldridge (2008). Discourse connective argument identification with connective specific rankers. *Proc. IEEE Conference on Semantic Computing (ICSC-08)*, Santa Clara CA.

[fos10] ○ George Foster, Pierre Isabelle, and Roland Kuhn (2010). Translating structured documents. *Proceedings of AMTA*, Denver CO.

[gal03] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing (2003). Discourse segmentation of multi-party conversation. *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics*.

[Gui12] Liane Guillou. Improving pronoun translation for statistical machine translation. *Proceedings, European Chapter of the Association for Computational Linguistics (EACL)*, 2012.

[guo] Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius (2010). Identifying the information structure of scientific abstracts. *Proc. 2010 BioNLP Workshop*, Uppsala, Sweden.

[Har93] Hardt, D. (1993). VP Ellipsis: Form, meaning, and processing. *PhD Dissertation. University of Pennsylvania.*

[hir08] Kenji Hirohata, Naoki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka (2008). Identifying sections in scientific abstracts using conditional random fields. *Proc. 3rd Int'l Joint Conf. on NLP*, 381–388.

[HBCb^{+}07] Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondřej Bojar. Moses: Open Source Toolkit for Statistical Machine Translation. pages 177–180, 2007.

[HF10] 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.

[hea97] Marti Hearst (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

[KN95] R. Kneser and H. Ney. Improved Backing-Off for M-gram Language Modeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:181–184, 1995.

[KO11] Moshe Koppel and Noam Ordan. Translationese and its dialects. *Proc. 49th Annual Meeting, Association for Computational Linguistics*, pages 1318–1326, 2011.

[lia10] Maria Liakata, Simone Teufel, Advaith Siddharthan and Colin Batchelor (2010). Corpora for the conceptualisation and zoning of scientific papers. *Proc.7th LREC*, Valletta, Malta.

[lin06] Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur (2006). Generative content models for structural analysis of medical abstracts. *Proc. HLT-NAACL Workshop on BioNLP*,

[Lop08] Adam Lopez. Statistical machine translation. *ACM Computing Surveys*, 40(3):49 pages, 2008.

[mal06] Igor Malioutov and Regina Barzilay (2006). Minimum cut model for spoken lecture segmentation. *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.
[marcu:02] Daniel Marcu and Abdessamad Echihabi (2002). An unsupervised approach to recognizing discourse relations. *Proceedings of ACL'02*.

[mckn03] Larry McKnight and Padmini Srinivasan (2003). Categorization of sentence types in medical abstracts. *Proc. AMIA Annual Symposium*, 440–444.

[mey11] Thomas Meyer (2011) Disambiguating temporal-contrastive connectives for machine translation. *Proc. ACL 2011 Student Session*, Portland, OR, 46–51,

[moe99] Marie-Francine Moens, Caroline Uyttendaele and Jos Dumortier (1999). Information extraction from legal texts: the potential of discourse analysis. *Int'l. Journal of Human-Computer Studies*, 51:1155–1171.

[moe00] Marie-Francine Moens, Caroline Uyttendaele and Jos Dumortier (2000). Intelligent information extraction from legal texts. *Information & Communications Technology Law*, 9:17–26.

[mt87] William Mann and Sandra Thompson (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

[Nie12] John Niekrasz (2012). Toward Summarization of Communicative Activities in Spoken Conversation. PhD thesis, University of Edinburgh.

[Nie05] Nielsen, L. (2004). Verb phrase ellipsis detection using automatically parsed text. *Proceedings of the 20th international conference on Computational Linguistics* (p. 1093).
[NK10] Ronan Le Nagard and Philipp Koehn. Aiding pronoun translation with co-reference resolution. *Proc. 5^{th} Joint Workshop on Statistical Machine Translation and Metrics (MATR)*, 2010.

[ON03] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51, March 2003.

[OGVW06] K Owczarzak, D Groves, J. van Genabith and A. Way (2006). Contextual bitext-derived paraphrases in automatic MT evaluation, *Proc., Workshop on Statistical Machine Translation*, pp. 86–93.

[pit09b] Emily Pitler and Ani Nenkova (2009). Using syntax to disambiguate explicit discourse connectives in text. *Proc. 47th Meeting of the Association for Computational Linguistics*.

[pra10] Rashmi Prasad, Aravind Joshi, and Bonnie Webber (2010). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. *Proc. COLING*, Beijing.

[PRWjZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu (2002). BLEU: a method for automatic evaluation of machine translation. *Proc. 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318.

[purv06] Matthew Purver, Konrad Körding, Thomas Griffiths and Joshua Tenenbaum (2006). Unsupervised Topic Modelling for Multi-Party Spoken Discourse. *Proc. 21^{st} COLING and 44^{th} Annual Meeting of the ACL*, Sydney, pp. 17–24.

[purv11] Matthew Purver (2011). Topic Segmentation. In Gokhan Tur and Renato de Mori (eds.), *Spoken Language Understanding*, Wiley.

[ruch07] Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti et al (2007). Using argumentation to extract key sentences from biomedical abstracts. *Int'l J. of Medical Informatics*, 76(2–3):195–200.

[Sto02] Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. *Proceedings of ICSLP*, v2, pp. 901–904, Denver CO, 2002.

[tab09] Maite Taboada, Julian Brooke and Manfred Stede (2009). Genre-based paragraph classification for sentiment analysis. *Proc. SIGDIAL 2009*, pp. 62–70.

[teu10] Simone Teufel (2010). *The Structure of Scientific Articles*. CSLI Publications, Stanford CA.

[wp07] Ben Wellner and James Pustejovsky (2007). Automatically identifying the arguments of discourse connectives. *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-07)*.