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

⇒ 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]
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: 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:
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 financing was 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.

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”.

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

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

Expository text can be viewed as a sequence of **topically coherent** segments. Their order may become conventionalized over time:

<table>
<thead>
<tr>
<th>Wisconsin</th>
<th>Louisiana</th>
<th>Vermont</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Etymology</td>
<td>2 History</td>
<td>3 Geography</td>
</tr>
<tr>
<td>4 Demographics</td>
<td>5 Law and government</td>
<td>6 Economy</td>
</tr>
<tr>
<td>7 Law and government</td>
<td>8 Education</td>
<td>9 Culture</td>
</tr>
<tr>
<td>10 ...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Wikipedia articles about US states

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,ma06].
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 (i.e., excluding "stop-words").

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].

Texts within a given genre – e.g.,

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

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

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 (i.e., 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.

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]

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.
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**

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?

---

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 syntactically bound to the connective
- **Arg1** – the other argument

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]

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.

<table>
<thead>
<tr>
<th>Type of connective</th>
<th>Ranking Accuracy</th>
<th>Re-ranking Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinating conjunctions</td>
<td>75.5%</td>
<td>78.3%</td>
</tr>
<tr>
<td>Subordinating conjunctions</td>
<td>87.2%</td>
<td>86.8%</td>
</tr>
<tr>
<td>Discourse adverbials</td>
<td>42.2%</td>
<td>49%</td>
</tr>
</tbody>
</table>

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

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

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.

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.

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:

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

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
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-occurrence features as predictors of the relation indicated by the clauses are conjoined.

---

**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.

---

**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”.

---

**Other 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.
- 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

- 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?
Disambiguating markers of discourse relations for SMT

Identifying clause-combining patterns for sentence-alignment in 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.

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

<table>
<thead>
<tr>
<th>Connective</th>
<th>Orig Frequency</th>
<th>Trans Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>therefore</td>
<td>0.153%</td>
<td>0.287%</td>
</tr>
<tr>
<td>nevertheless</td>
<td>0.019%</td>
<td>0.045%</td>
</tr>
<tr>
<td>thus</td>
<td>0.015%</td>
<td>0.041%</td>
</tr>
<tr>
<td>moreover</td>
<td>0.008%</td>
<td>0.035%</td>
</tr>
</tbody>
</table>

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

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.


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.
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


both using effectively the same procedure.

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.

Earlier results showed a small but disappointing improvement:

  
  - 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.

3rd-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

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 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.

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.
How well does Google do?

VPE with a subject pronoun:

*He doesn’t want to speak, so she will.*

*Ils 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.*

*Ils 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 ≥50% use a common reduced form:

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Example</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJ+BE IT</td>
<td>Certaines sont bonnes et certaines ne le sont pas.</td>
<td>58</td>
</tr>
<tr>
<td>SUBJ+DO IT</td>
<td>Je pourrais citer des pays, je ne le ferai pas.</td>
<td>47</td>
</tr>
<tr>
<td>IT+BE THE CASE</td>
<td>Vous auriez pu r'égler tout ceci mais cela n'a pas été le cas.</td>
<td>39</td>
</tr>
<tr>
<td>SUBJ + NOT</td>
<td>Nous avons le temps, Saddam pas.</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
- 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 \([\text{they}] \text{ do not}\)

Some countries ban organisations, others \((\text{they}) \text{ do not}\).
Animals have rights, children \((\text{they}) \text{ do not}\).
Those large ones employ staff and the small ones \((\text{they}) \text{ do not}\).  

Possible translations:
- Certains états membres pratiquent un système de cartes, d’autres non.
- Certains états membres pratiquent un système de cartes, d’autres pas.
- Certains états membres 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


- 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 peuvent 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


Introduction
Aspects of discourse relevant to SMT
How discourse can contribute to SMT
Conclusion


