STS for Machine Translation Evaluation

STS Workshop, NYC

March 12-13 2012

Lucia Specia

University of Sheffield
l.specia@sheffield.ac.uk

Outline

1. Monolingual STS
   - MT Evaluation against references
   - TINE
2. Multilingual STS
   - MT Evaluation without references
   - Adequacy estimation - assimilation purposes
3. STS for Evaluation
   - One metric fits evaluation for all applications?
   - One metric fits all applications?
4. My 2 cents
   - STS from an application perspective

Monolingual STS

- Meteor - inexact lexical/phrase matching

Monolingual STS

- Meteor - inexact lexical/phrase matching
- Pado et al. - textual entailment features

Monolingual STS

- Meteor - inexact lexical/phrase matching
- Pado et al. - textual entailment features
- Gimenez & Marquez - matching of semantic labels

Monolingual STS

- Meteor - inexact lexical/phrase matching
- Pado et al. - textual entailment features
- Gimenez & Marquez - matching of semantic labels
- Meant - matching of semantic roles (predicates and their arguments)

Monolingual STS

- Meteor - inexact lexical/phrase matching
- Pado et al. - textual entailment features
- Gimenez & Marquez - matching of semantic labels
- Meant - matching of semantic roles (predicates and their arguments)
- TINE - matching of **semantic roles** (predicates and their arguments), but **automatically**

Tine Is Not Entailment

R: The lack of snow is **putting** [people]$_{A0}$ **off** **booking** [ski holidays]$_{A1}$ in [hotels and guest houses]$_{AM-LOC}$.

H: The lack of snow **discourages** [people]$_{A0}$ from **ordering** [ski stays]$_{A1}$ in [hotels and boarding houses]$_{AM-LOC}$.

Tine Is Not Entailment

R: The lack of snow is **putting** [people]$_{A0}$ **off** **booking** [ski holidays]$_{A1}$ in [hotels and guest houses]$_{AM-LOC}$.

H: The lack of snow **discourages** [people]$_{A0}$ from **ordering** [ski stays]$_{A1}$ in [hotels and boarding houses]$_{AM-LOC}$.

Lexical matching component L & **semantic** component A:

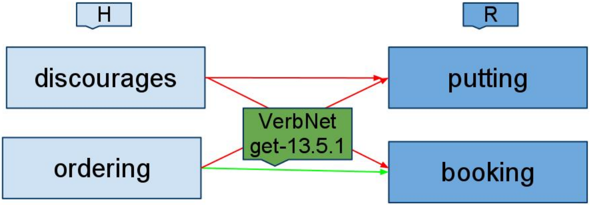
$$T(H, \mathbf{R}) = max \left\{ \frac{\alpha L(H,R) + \beta A(H,R)}{\alpha + \beta} \right\}_{R \in \mathbf{R}}$$

This Is Not Entailment

L: **BLEU**; **S**: **matching** of **verbs** and their **arguments**:

$$A(H,R) = \frac{\sum_{v \in V} verb\_score(H_v, R_v)}{|V_r|}$$

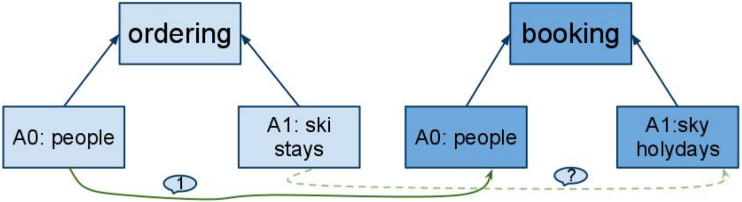
1. Align verbs using ontologies (VerbNet and VerbOcean):



v_h and v_r are aligned if they share a class in **VerbNet** or hold a relation in **VerbOcean**

This Is Not Entailment

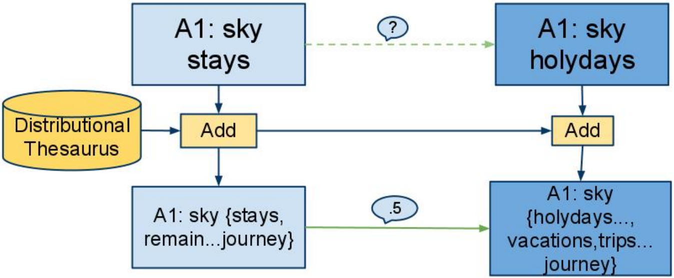
2. Match arguments with same semantic roles:



$$verb\_score(H_v, R_v) = \frac{\sum_{a \in A_h \cap A_r} arg\_score(H_a, R_a)}{|A_r|}$$

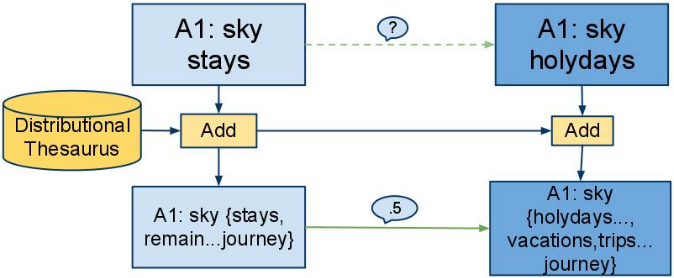
This Is Not Entailment

3. Expand arguments using distributional semantics and match them using cosine similarity: $arg\_score(H_a, R_a)$



This Is Not Entailment

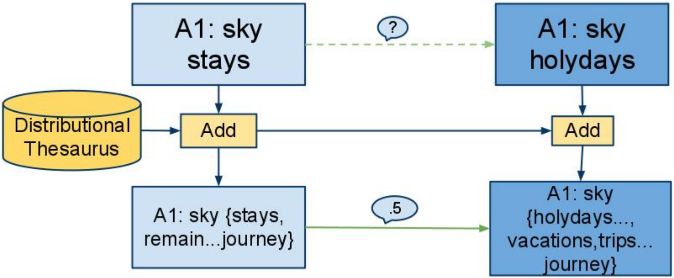
3. Expand arguments using distributional semantics and match them using cosine similarity: $arg\_score(H_a, R_a)$



TINE did slightly better than BLEU at **segment level**.

This Is Not Entailment

3. Expand arguments using distributional semantics and match them using cosine similarity: $arg\_score(H_a, R_a)$



TINE did slightly better than BLEU at **segment level**.
Lexical component extremely important.

Quality Estimation

- No access to reference translation - **MT system in use**: post-editing, dissemination, assimilation, etc

Quality Estimation

- No access to reference translation - **MT system in use**: post-editing, dissemination, assimilation, etc
- Semantics particularly important for estimating **adequacy**

Quality Estimation

- No access to reference translation - **MT system in use**: post-editing, dissemination, assimilation, etc
- Semantics particularly important for estimating **adequacy**

Example 1

Target:
Chang-e III is expected to launch **after** 2013

Source:
嫦娥三号预计 2013 年前后发射

Reference:
Chang-e III is expected to launch **around** 2013

By Google Translate

Example 2

Target:
Continued high floods **subside**. Guang'an old city has been soaked 2 days 2 nights

Source:
四川广安洪水持续高位不退 老城区已被泡 2 天 2 夜

Reference:
The continuing floods in Guang'an - Sichuan have **not subsided**. The old city has been flooded for 2 days and 2 nights.

By Google Translate

Example 3

Target:
site security should be included in **sex education** curriculum for students

Source:
场地安全性教育应纳入学生的课程

Reference:
site security **requirements** should be included in the **education** curriculum for students

By Google Translate

Most common problems

- words translated incorrectly
- incorrect relationship: words/constituents/clauses
- missing/untranslated/repeated/added words
- incorrect word order
- inflectional/voice error

MT quality evaluation

- How does the metrics vary depending on how the **references** are produced?
  - Standard references - semantic component only, segment-level correlation: 0.21
  - Post-edited translations - semantic component only, segment-level correlation: 0.55

MT quality evaluation vs intrinsic evaluation

- TINE on WMT data: correlation: 0.30
- TINE on Microsoft video data: correlation: 0.43
- TINE on Microsoft paraphrase data: correlation: 0.30

MT quality estimation and evaluation

- Can we use the same approach as **reference-based evaluation**, but **bilingual**?

MT quality estimation and evaluation

- Can we use the same approach as **reference-based evaluation**, but **bilingual**?
  - Possibly, assuming resources and alignments are available

MT quality estimation and evaluation

- Can we use the same approach as **reference-based evaluation**, but **bilingual**?
  - Possibly, assuming resources and alignments are available
  - Cannot expect exact correspondences. E.g. thematic divergences (Dorr et al):

MT quality estimation and evaluation

- Can we use the same approach as **reference-based evaluation**, but **bilingual**?
  - Possibly, assuming resources and alignments are available
  - Cannot expect exact correspondences. E.g. thematic divergences (Dorr et al):

    I miss you vs. **Tu me manques**

MT quality estimation and evaluation

- Can we use the same approach as **reference-based evaluation**, but **bilingual**?
  - Possibly, assuming resources and alignments are available
  - Cannot expect exact correspondences. E.g. thematic divergences (Dorr et al):

    I miss you vs. **Tu me manques**
  - Can learn these correspondences

MT evaluation and summarization (evaluation)

- Can the same STS metric address both?

MT evaluation and summarization (evaluation)

- Can the same STS metric address both?
  - MT systems make **mistakes** that summarization (esp. extractive) systems are not likely to make

MT evaluation and summarization (evaluation)

- Can the same STS metric address both?
  - MT systems make **mistakes** that summarization (esp. extractive) systems are not likely to make
  - Translation is generally **related/similar** to the reference (and source) (a 1-2 likert score), not the case in summarization

MT evaluation and summarization (evaluation)

- Can the same STS metric address both?
  - MT systems make **mistakes** that summarization (esp. extractive) systems are not likely to make
  - Translation is generally **related/similar** to the reference (and source) (a 1-2 likert score), not the case in summarization
  - Translation is generally very similar in **length** to the reference, not the case in summarization

MT evaluation and summarization (evaluation)

- Can the same STS metric address both?
  - MT systems make **mistakes** that summarization (esp. extractive) systems are not likely to make
  - Translation is generally **related/similar** to the reference (and source) (a 1-2 likert score), not the case in summarization
  - Translation is generally very similar in **length** to the reference, not the case in summarization
  - Translation does **1-1 comparisons**, not the case in summarization

MT evaluation and summarization (evaluation)

- Can the same STS metric address both?
  - MT systems make **mistakes** that summarization (esp. extractive) systems are not likely to make
  - Translation is generally **related/similar** to the reference (and source) (a 1-2 likert score), not the case in summarization
  - Translation is generally very similar in **length** to the reference, not the case in summarization
  - Translation does **1-1 comparisons**, not the case in summarization

Translation needs a more **fine-grained** metric than summarization?

Applications require different STS metrics

- "How do we illustrate the utility of STS to end applications?"

Applications require different STS metrics

- "How do we illustrate the utility of STS to end applications?"
- STS depends on what is **important** for the application, and also on the sort of **data** that can be produced by them

Applications require different STS metrics

- "How do we illustrate the utility of STS to end applications?"
- STS depends on what is **important** for the application, and also on the sort of **data** that can be produced by them
- Avoid falling into the same trap as **WSD**?

Applications require different STS metrics

- "How do we illustrate the utility of STS to end applications?"
- STS depends on what is **important** for the application, and also on the sort of **data** that can be produced by them
- Avoid falling into the same trap as **WSD**?
- How many applications use an **off-the-shelf WSD module**?

Applications require different STS metrics

- "How do we illustrate the utility of STS to end applications?"
- STS depends on what is **important** for the application, and also on the sort of **data** that can be produced by them
- Avoid falling into the same trap as **WSD**?
- How many applications use an **off-the-shelf WSD module**?
- **Common excuses**: not good enough, WN senses not appropriate for my application...

Applications require different STS metrics

- "How do we illustrate the utility of STS to end applications?"
- STS depends on what is **important** for the application, and also on the sort of **data** that can be produced by them
- Avoid falling into the same trap as **WSD**?
- How many applications use an **off-the-shelf WSD module**?
- **Common excuses**: not good enough, WN senses not appropriate for my application...

STS from an application perspective

- Select a few **applications** that could benefit from STS

STS from an application perspective

- Select a few **applications** that could benefit from STS
- Collect examples with different levels of similarity **for these applications**

STS from an application perspective

- Select a few **applications** that could benefit from STS
- Collect examples with different levels of similarity **for these applications**
- **Gold-standard** annotation for these examples (like in Meant)

STS from an application perspective

- Select a few **applications** that could benefit from STS
- Collect examples with different levels of similarity **for these applications**
- **Gold-standard** annotation for these examples (like in Meant)
- Compute as many **semantic components** as possible (word-level, SRL, etc)

STS from an application perspective

- Select a few **applications** that could benefit from STS
- Collect examples with different levels of similarity **for these applications**
- **Gold-standard** annotation for these examples (like in Meant)
- Compute as many **semantic components** as possible (word-level, SRL, etc)
- I'm not sure components need to talk to each other: **error propagation**
- Regress on these to understand what are the **important components** for each application

STS from an application perspective

- Select a few **applications** that could benefit from STS
- Collect examples with different levels of similarity **for these applications**
- **Gold-standard** annotation for these examples (like in Meant)
- Compute as many **semantic components** as possible (word-level, SRL, etc)
- I'm not sure components need to talk to each other: **error propagation**
- Regress on these to understand what are the **important components** for each application
- Repeat process with **automatic annotation**

STS from an application perspective

- Select a few **applications** that could benefit from STS
- Collect examples with different levels of similarity **for these applications**
- **Gold-standard** annotation for these examples (like in Meant)
- Compute as many **semantic components** as possible (word-level, SRL, etc)
- I'm not sure components need to talk to each other: **error propagation**
- Regress on these to understand what are the **important components** for each application
- Repeat process with **automatic annotation**

Parameterizable STS metric

STS for Machine Translation Evaluation

STS Workshop, NYC

March 12-13 2012

Lucia Specia

University of Sheffield
l.specia@sheffield.ac.uk