Bilingual MRS Alignment

Michael Wayne Goodman

Nanyang Technological University
& University of Washington

goodmami@uw.edu

Overview

1. Background
2. Representations and Useful Information
3. Alignment Methods

Background

- Last year I began my research into MT by statistical transfer
- This project involves at least two components:
  1. Training a statistical transfer model
  2. Decoding input MRS using the model
- Today I discuss some findings from the **training** component

Recall: why use semantics?

- Consider context sensitivity:
  1. 犬が吠える | inu-ga hoeru | *The dog barks*
  2. ライオンが吠える | raion-ga hoeru | *The lion roars*

- 犬 --> *dog*
- ライオン --> *lion*
- 吠える --> *bark* (for dog); *roar* (for lion); etc..

- Or idioms: "to kick the bucket" -> "to die"
- passivization (was chased)
- long distance dependencies (the lion who I thought my mother heard roar yesterday) 昨日母が吠えることが聞こえたと私が思ったライオン

Plug: pyDelphin

- Mathieu Morey began work with graph matching of MRS
- Inspired by his work, I've refactored pyDelphin to use an actual graph representation underlying its API for MRS and DMRS methods
- This has made working with MRSs in this context more effective
- Hopefully I've learned from the challenges Mathieu came across, although I'm not currently doing the graph-matching he was

Plug: pyDelphin

- Mathieu Morey began work with graph matching of MRS
- Inspired by his work, I've refactored pyDelphin to use an actual graph representation underlying its API for MRS and DMRS methods
- This has made working with MRSs in this context more effective
- Hopefully I've learned from the challenges Mathieu came across, although I'm not currently doing the graph-matching he was

`https://github.com/goodmami/pydelphin`

Representations and Useful Information

Representations and useful information

- Serialization of MRS is mostly orthogonal to issues of transfer
- But the structure of that information is useful

- E.g. Transfer should not care about:
  - Variables (but variable sort?)
  - Labels (but quantifier scope?)
  - Lnk (CFROM and CTO)
  - Node identifiers

- Transfer should consider:
  - predicates
  - how predicates relate to each other
  - maybe the top pred/scope

Representations and useful information

Consider these two sentences:

- "I am sure that I shall say nothing of the kind."
- "いやいや、そんなことは言わんよ。"

MRS

```
[ LTOP: h0
  INDEX: e2
  RELS: < [ pron_rel<0:1> LBL: h4 ARG0: x3  ]
          [ pronoun_q_rel<0:1> LBL: h5 ARG0: x3 RSTR: h6 BODY: h7 ]
          [ "_sure_a_of_rel"<5:9> LBL: h1 ARG0: e2 ARG1: x3 ARG2: h8 ]
          [ pron_rel<15:16> LBL: h9 ARG0: x10 ]
          [ pronoun_q_rel<15:16> LBL: h11 ARG0: x10 RSTR: h12 BODY: h13 ]
          [ "_say_v_1_rel"<23:26> LBL: h14 ARG0: e15  ARG1: x10 ARG2: x16 ]
          [ thing_rel<27:34> LBL: h17 ARG0: x16 ]
          [ _no_q_rel<27:34> LBL: h18 ARG0: x16 RSTR: h19 BODY: h20 ]
          [ _of_p_rel<35:37> LBL: h17 ARG0: e21  ARG1: x16 ARG2: x22 ]
          [ _the_q_rel<38:41> LBL: h23 ARG0: x22 RSTR: h24 BODY: h25 ]
          [ "_kind_n_of-n_rel"<42:47> LBL: h26 ARG0: x22 ARG1: i27 ] >
  HCONS: < h0 qeq h1 h6 qeq h4 h8 qeq h14 h12 qeq h9 h19 qeq h17 h24 qeq h26 > ]
```

```
[ LTOP: h0
  INDEX: e2
  RELS: < [ "_iyaiya_a_2_rel"<0:4> LBL: h1 ARG0: e4  ARG1: e2 ]
          [ "_sono_q_rel"<7:10> LBL: h5 ARG0: x6 RSTR: h7 BODY: h8 ]
          [ "_koto_n_nom_rel"<11:13> LBL: h9 ARG0: x6 ]
          [ _wa_d_rel<14:15> LBL: h9 ARG0: e10  ARG1: e11  ARG2: x6 ]
          [ "_iu_v_2_rel"<16:18> LBL: h12 ARG0: e2 ARG1: i13 ARG2: x6 ]
          [ "_neg_v_rel"<19:21> LBL: h1 ARG0: i14 ARG1: h15 ] >
  HCONS: < h0 qeq h1 h7 qeq h9 h15 qeq h12 > ]
```

DMRS

```
dmrs {
  10000 [pron_rel<0:1> x];
  10001 [pronoun_q_rel<0:1> x];
  10002 ["_sure_a_of_rel"<5:9> e];
  10003 [pron_rel<15:16> x];
  10004 [pronoun_q_rel<15:16> x];
  10005 ["_say_v_1_rel"<23:26> e];
  10006 [thing_rel<27:34> x];
  10007 [_no_q_rel<27:34> x];
  10008 [_of_p_rel<35:37> e];
  10009 [_the_q_rel<38:41> x];
  10010 ["_kind_n_of-n_rel"<42:47> x];
  0:/H -> 10002;
  10001:RSTR/H -> 10000;
  10002:ARG1/NEQ -> 10000;
  10002:ARG2/H -> 10005;
  10004:RSTR/H -> 10003;
  10005:ARG1/NEQ -> 10003;
  10005:ARG2/NEQ -> 10006;
  10007:RSTR/H -> 10006;
  10008:ARG1/EQ -> 10006;
  10008:ARG2/NEQ -> 10010;
  10009:RSTR/H -> 10010;
}
```

```
dmrs {
  10000 ["_iyaiya_a_2_rel"<0:4> e];
  10001 ["_sono_q_rel"<7:10> x];
  10002 ["_koto_n_nom_rel"<11:13> x];
  10003 [_wa_d_rel<14:15> e];
  10004 ["_iu_v_2_rel"<16:18> e];
  10005 ["_neg_v_rel"<19:21> i];
  0:/H -> 10000;
  10000:ARG1/NEQ -> 10004;
  10000:/EQ -- 10005;
  10001:RSTR/H -> 10002;
  10003:ARG2/EQ -> 10002;
  10004:ARG2/NEQ -> 10002;
  10005:ARG1/H -> 10004;
}
```

Elementary Dependency Structures

```
{e3:
 x5:pron⟨0:1⟩[]
 _1:pronoun_q⟨0:1⟩[BV x5]
 e3:_sure_a_of⟨5:9⟩[ARG1 x5, ARG2 e16]
 x11:pron⟨15:16⟩[]
 _2:pronoun_q⟨15:16⟩[BV x11]
 e16:_say_v_1⟨23:26⟩[ARG1 x11, ARG2 x17]
 x17:thing⟨27:34⟩[]
 _3:_no_q⟨27:34⟩[BV x17]
 e22:_of_p⟨35:37⟩[ARG1 x17, ARG2 x23]
 _4:_the_q⟨38:41⟩[BV x23]
 x23:_kind_n_of-n⟨42:47⟩[]
}
```

```
{e2:
 e4:_iyaiya_a_2(0:4)[ARG1 e2]
 _1:_sono_q(7:10)[BV x6]
 x6:_koto_n_nom(11:13)[]
 e10:_wa_d(14:15)[ARG2 x6]
 e2:_iu_v_2(16:18)[ARG2 x6]
 i14:_neg_v(19:21)[ARG1 e2]
}
```

MRS-Path

- MRS-Path is tree (non-reentrant) of DMRS subgraphs
- It cannot always represent a full MRS (at least, not without composition)
- It may have ambiguous nodes (two of the same pred in one path)

E.g.

```
_sure_a_of(:ARG1/NEQ>pron :ARG2/H>_say_v_1(:ARG1/NEQ>pron :ARG2/NEQ>thing))
_sure_a_of(:ARG1/NEQ>pron :ARG2/H>_say_v_1(:ARG1/NEQ> :ARG2/NEQ>))
_sure_a_of(:ARG1/NEQ>pron :ARG2/H>_say_v_1)
_sure_a_of(:ARG1/NEQ> :ARG2/H>)
_say_v_1(:ARG1/NEQ>pron :ARG2/NEQ>thing)
_of_p(:ARG1/NEQ>thing :ARG2/NEQ>_kind_n_of-n)
...
```

```
_iyaiya_a_2:ARG1/NEQ>_iu_v_2(:ARG2/NEQ>_koto_n_nom)
_neg_v:ARG1/H>_iu_v_2(:ARG2/NEQ>_koto_n_nom)
_iyaiya_a_2:/EQ>_neg_v
_iu_v_2(:ARG2/NEQ>_koto_n_nom)
_iu_v_2(:ARG1/NEQ> :ARG2/NEQ>_koto_n_nom)
_sono_q:RSTR/H>_koto_n_nom
_wa_d:ARG2/EQ>_koto_n_nom
...
```

Alignment Methods

Pred-only replacement

- Training
  - Extract linearization of predicates from MRS
  - Align predicates with off-the-shelf word aligners (1-to-1)
- Decoding
  - Replace preds with best-match (unigram) alignment
  - Preserve MRS structure (EPs (modulo preds), args, LTOP, etc)

Pred-only replacement

- Training
  - Extract linearization of predicates from MRS
  - Align predicates with off-the-shelf word aligners (1-to-1)
- Decoding
  - Replace preds with best-match (unigram) alignment
  - Preserve MRS structure (EPs (modulo preds), args, LTOP, etc)
- Abysmal results, as expected
  - quantifier mismatch (udef_q vs pronoun_q)
  - incompatible properties
  - fixing these got 24/127 items from the MRS testsuite to generate

```
This tarou 2 barked.
This window was open.
This tarou 2 followed the jirou 2.
This tarou 2 followed the jirou 2.
This cat barked.
This cat followed the dog.
It barked.
This cat followed itself.
This cat followed himself.
...
```

Direct alignment of sub-graphs

- Training
  - Enumerate subgraphs for each MRS as MRS-Paths
  - Pass these into an aligner directly
  - treat subgraphs/MRS-Paths as tokens
  - 1-to-1 (or ngrams?)
- Decoding
  - TBD

Some alignments look nice:

```
_kau_v_1:ARG2/NEQ>    _buy_v_1:ARG2/NEQ>
_kono_q:RSTR/H>_hon_n    _this_q_dem:RSTR/H>_book_n_of
_sha_a_4:ARG1/EQ>_kagaku_n_1    _scientist_n_1
nominalization:ARG1/HEQ_doryoku_s    _effort_n_1
```

Others not so much:

```
_kuru_v    _come_v_1:ARG1/NEQ>pron
_kaku_v_10:ARG2/NEQ>_tegami_n    _write_v_to:ARG1/NEQ>named
```

Graph-matching by proxy of pred alignment

- Training
  - Linearize and align preds as in the first method (but as n-grams)
  - Use n-grams of aligned preds to select **connected** MRS subgraphs
  - Similar to Petter Haugereid's work, but without templates
- Decoding
  - TBD (but same as previous method)

Quantifiers

- Quantifiers complicate the graph
- …but are necessary for a well-formed MRS
- Do we even want them in the model?
- Consider, "Every day" vs "毎日"

```
dmrs {
  10004 [_every_q_rel<10:15> x PERS=3 NUM=sg IND=+];
  10005 ["_day_n_of_rel"<16:20> x PERS=3 NUM=sg IND=+];
  10004:RSTR/H -> 10005;
}
```

```
dmrs {
  10001 [udef_q_rel<0:2> x ];
  10002 ["_mainichi_n_rel"<0:2> x ];
  10001:RSTR/H -> 10002;
}
```

Future work: Collapsing structures (1)

- Some semantic constructs create graph distance between more contentful nodes
- Graph distance can make it more difficult to automatically discover useful alignments

- Prepositions: e.g. "fly to Lisbon", instead of

  ```
  to_p(:ARG1/NEQ>_fly_v_1 :ARG2/NEQ>"Lisbon")
  ```

  We get:

  ```
  _fly_v_1:<to_p>"Lisbon"
  ```

Future work: Collapsing structures (2)

- Coordination: e.g. for "dogs and cats sleep", instead of

  ```
  _sleep_v_1:ARG1/NEQ>_and_c(:L-INDEX/NEQ>_dog_n_1 :R-INDEX/NEQ>_cat_n_1)
  ```

  We get:

  ```
  _sleep_v_1:ARG1/NEQ>(_dog_n_1 & _cat_n_1)
  ```

Future work: Native graph-based alignment

- Adapt existing alignment methods (e.g. Giza++, Anymalign) to work on graphs instead of ngraphs or trees
- The alignment can be tuned by what we know about well-formed MRS
- Subgraphs can be selected by graph structure rather than something approximate (like CFROM-aligned preds)

Other Future Work:

- Integrating WordNet
- qMRS

References

- Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A. Sag. "Minimal recursion semantics: An introduction." Research on Language and Computation 3, no. 2-3 (2005): 281-332.
- Copestake, Ann. "Robust minimal recursion semantics." unpublished draft (2006).
- Copestake, Ann. "Slacker semantics: why superficiality, dependency and avoidance of commitment can be the right way to go." In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1-9. Association for Computational Linguistics, 2009.
- Haugereid, Petter, and Francis Bond. "Extracting semantic transfer rules from parallel corpora with SMT phrase aligners." Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation. Association for Computational Linguistics, 2012.
- Oepen, Stephan, and Jan Tore Lønning. "Discriminant-based MRS banking." In Proceedings of the 5th international conference on language resources and evaluation (lrec 2006). 2006.