A Discriminative Model for Semantics-to-String Translation

Aleš Tamchyna[1] and Chris Quirk[2] and Michel Galley[2]

[1]Charles University in Prague [2]Microsoft Research

July 30, 2015

Introduction

- State-of-the-art MT models still use a simplistic view of the data
  - words typically treated as independent, unrelated units
  - relations between words only captured through linear context
- Unified semantic representations, such as Abstract Meaning Representation (AMR, Banarescu et al. 2013), (re)gaining popularity
- Abstraction from surface words, semantic relations made explicit, related words brought together (possibly distant in the surface realization)
- Possible uses:
  - Richer models of source context ← **our work**
  - Target-side (or joint) models to capture semantic coherence
  - Semantic transfer followed by target-side generation

Semantic Representation

- Logical Form transformed into an AMR-style representation (Vanderwende et al., 2015)
- Labeled directed graph, not necessarily acyclic (e.g. coreference)
- Nodes ∼ content words, edges ∼ semantic relations
- Function words (mostly) not represented as nodes
- “Bits” capture various linguistic properties

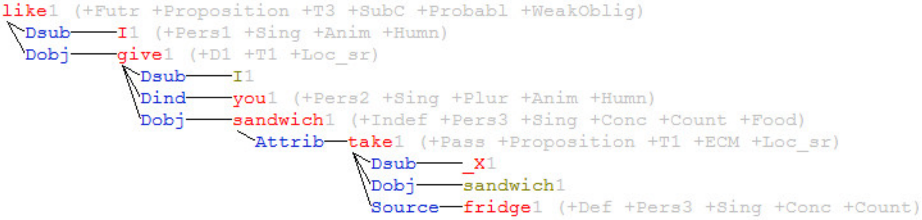


Figure 1 : Logical Form (computed tree) for the sentence: *I would like to give you a sandwich taken from the fridge.*

Graph-to-String Translation

Translation = generation of target-side surface words in order, conditioned on source semantic nodes and previously generated words.

- Start in the (virtual) root
- At each step, transition to a semantic node and emit a target word
- A single node can be visited multiple times
- One transition can move anywhere in the LF

Source-side semantic graph: $G = (V, E)$, $V = \{n_1, ..., n_S\}$, $E \subset V \times V$
Target string $E = (e_1, ..., e_T)$, alignment $A = (a_1, ..., a_T)$, $a_i \in 0...S$.

$$P(A, E|G) = \prod_{i=1}^{T} P(a_i|a_1^{i-1}, e_1^{i-1}, G) P(e_i|a_1^{i}, e_1^{i-1}, G)$$

Translation Example

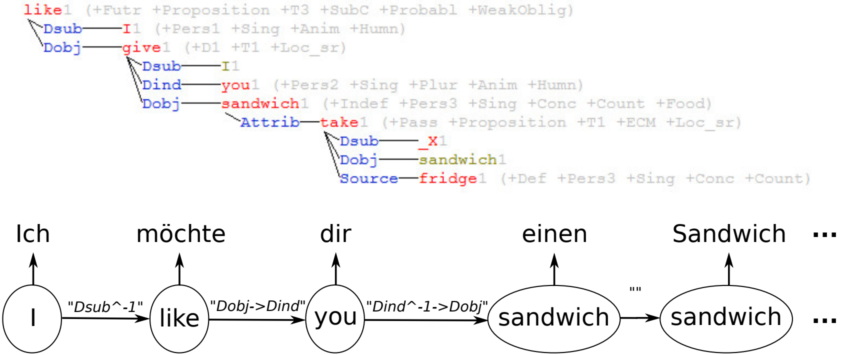


Figure 2 : An example of the translation process illustrating several first steps of translating the sentence into German (“*Ich möchte dir einen Sandwich...*”). Labels in italics correspond to the shortest undirected paths between the nodes.

Alignment of Graph Nodes

How do we align source-side semantic nodes to target-side words?

Evaluated approaches:

1. Gibbs sampling
2. Direct GIZA++
3. Alignment composition

Alignment of Graph Nodes – Gibbs Sampling

Alignment ($\sim$ transition) distribution $P(a_i|\cdots)$ modeled as a categorical distribution:

$$P(a_i|a_{i-1}, G) \propto \mathrm{c}(\textsc{label}(a_{i-1}, a_i))$$

Translation ($\sim$ emission) distribution modeled as a set of categorical distributions, one for each source semantic node:

$$P(e_i|n_{a_i}) \propto \mathrm{c}(\textsc{lemma}(n_{a_i}) \rightarrow e_i)$$

Sample from the following distribution:

$$\begin{aligned} P(t|n_i) \propto & \frac{\mathrm{c}(\textsc{lemma}(n_i) \rightarrow t) + \alpha}{\mathrm{c}(\textsc{lemma}(n_i)) + \alpha L} \\ & \times \frac{\mathrm{c}(\textsc{label}(n_i, n_{i-1})) + \beta}{T + \beta P} \\ & \times \frac{\mathrm{c}(\textsc{label}(n_{i+1}, n_i)) + \beta}{T + \beta P} \end{aligned}$$

Alignment of Graph Nodes – Evaluation

2. Direct GIZA++
    - Linearize the LF, run GIZA++ (standard word alignment)
    - Heuristic linearization, try to preserve source surface word order
3. Alignment composition
    - Source-side nodes to source-side tokens
        - Parser-provided alignment
        - GIZA++
    - Source-target word alignment – GIZA++

Manual inspection of alignments

- Alignment composition clearly superior
- Not much difference between GIZA++ and parser alignments

Discriminative Translation Model

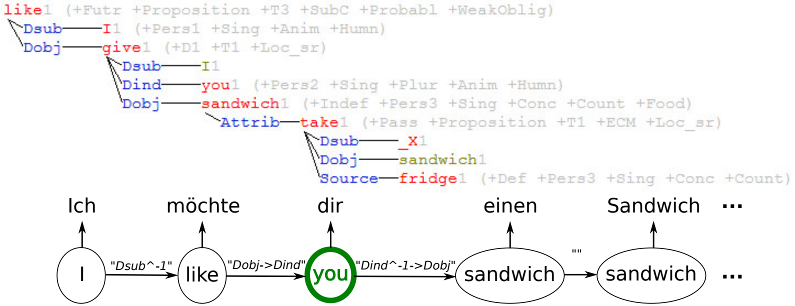
- A maximum-entropy classifier

$$P(e_i | n_{a_i}, n_{a_{i-1}}, G, e_{i-k+1}^{i-1}) = \frac{\exp\left(\vec{w} \cdot \vec{f}(e_i, n_{a_i}, n_{a_{i-1}}, G, e_{i-k+1}^{i-1})\right)}{Z}$$

$$Z = \sum_{e' \in GEN(n_{a_i})} \exp(\vec{w} \cdot \vec{f}(e', n_{a_i}, n_{a_{i-1}}, G, e_{i-k+1}^{i-1}))$$

- Possible classes: top 50 translations observed with given lemma
- Online learning with stochastic gradient descent
- Learning rate 0.05, cumulative L1 regularization with weight 1, batch size 1, 22 hash bits
- Early stopping when held-out perplexity increases
- Parallelized (multi-threading) and distributed learning for tractability

Feature Set



- **Current node, previous node, parent node** – lemma, POS, bits
- **Path from previous node** – path length, path description
- **Bag of lemmas** – capture overall topic of the sentence
- **Graph context** – features from nodes close in the graph (limited by the length of shortest undirected path)
- **Generated tokens** – “fertility”; some nodes should generate a function word first (e.g. an article) and then the content word
- **Previous tokens** – target-side context

Experiments

- Evaluated in a *n*-best re-ranking experiment
  - Generate 1000-best translations of devset sentences
  - Add scores from our model
  - Re-run MERT on the enriched *n*-best lists
- Basic phrase-based system, French→English
- 1 million parallel training sentences
- Obtained small but consistent improvements
- Differences would most likely be larger after integration in decoding

Dataset	Baseline	+Semantics
WMT 2009 = devset	17.44	17.55
WMT 2010	17.59	17.64
WMT 2013	17.41	17.55

Table 1 : BLEU scores of *n*-best reranking in French→English translation.

Conclusion

- Initial attempt at including semantic features in statistical MT
- Feature set comprising morphological, syntactic and semantic properties
- Small but consistent improvement of BLEU

Future work:

- Integrate directly in the decoder
- Parser accuracy limited – use multiple analyses
- Explore other ways of integration
  - Target-side models of semantic plausibility
  - Semantic transfer and generation

Thank You!

Questions?

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-2322`.

Lucy Vanderwende, Arul Menezes, and Chris Quirk. An AMR parser for English, French, German, Spanish and Japanese and new AMR-annotated corpus. In *Proceedings of the 2015 NAACL HLT Demonstration Session*, Denver, Colorado, June 2015. Association for Computational Linguistics.