Hybrid Interlingual Machine Translation

University of Malta
21 March 2013

Aarne Ranta
University of Gothenburg

Plan

Talk: introducing the hybrid model

Demo: some initial work for German, Hindi, Finnish, Maltese

Lecture: translation in GF - some more details and pointers

Joint work with Krasimir Angelov, John Camilleri, Ramona Enache, Shafqat Virk

Talk

Waves of Machine Translation

1948 Shannon, Weaver: statistical

1966 ALPAC: linguistic

1990 IBM: statistical

201? Church, Koehn: hybrid?

The virtues of SMT

Robustness: deals with any input (by smoothing)

Low cost: "just data", no language expertise

- Google translate's 60 languages

Captures common idioms

Problems with SMT: morphology

"Pure" SMT: every form is a separate word. Some of them are probably not seen in training data.

Finnish:

yö, yön, yötä, yönä, yöksi, yössä, yöstä, yöhön, yöllä, yöltä, yölle, yöttä, öineen, öin, yöt, öitä, öiden, öinä, öiksi, öissä, öistä, öihin, öillä, öiltä, öille, öittä, öin

English:

Night, night, night, night, night, night, night, night, night, night, nights, yöttä, öineen, night, night, nights, nights, nights States by quotas, domestic insurance companies, nights, nights, öillä, against loss, States, öittä, night

Problems with SMT: long-distance dependencies

Translation degrades as distance grows

*Er **bringt** dich **um**.*	*He will kill you.*
*Er **bringt** deinen Freund **um**.*	*He'll kill your friend.*
*Er **bringt** deinen ältesten Freund **um**.*	*He brings to your oldest friend.*

Sparseness of data

The common factor of both problems

- Finnish has 100k * 1000 word forms
- German needs n-grams as long as it has sentences

Factored models

To help sparseness of data:

SMT on lemma+tag pairs rather than strings

Adds an element of **linguistic knowledge**

What is data

Nothing says that data must be raw data (strings).

Data, in general, is **data structures**!

Lists, records, trees,...

Declarativity

One more virtue of SMT: it is based on **declarative language models**

- as opposed to procedural translation rules (**transfer**).

The fundamental equation:

$$\hat{e} = argmax\ P(f|e)P(e)$$

The statistical translation model

$$\hat{e} = argmax\ P(f|e)P(e)$$

To translate *f*, find the *e* that gives the best product of

- *P(f|e)*, probability of source *f* yielding *e*
- *P(e)*, probability of *e* in the target language

In "pure" SMT, *f* and *e* are strings of words.

But nothing says they must be!

Translating trees

$$\hat{e} = argmax\ P(f|e)P(e)$$

To translate the **tree** *f*, find the **tree** *e* that gives the best product of

- *P(f|e)*, probability of source **tree** *f* yielding target **tree** *e*
- *P(e)*, probability of **tree** *e* in the target language

What trees

Parse trees

- language-specific
- sensitive to word order
- leaves are word forms

Abstract syntax trees

- language-neutral
- pure constituency
- leaves are abstract lemmas ("senses")

How does this help?

Morphology:

- trees abstract away from word inflection

Long-distance dependencies:

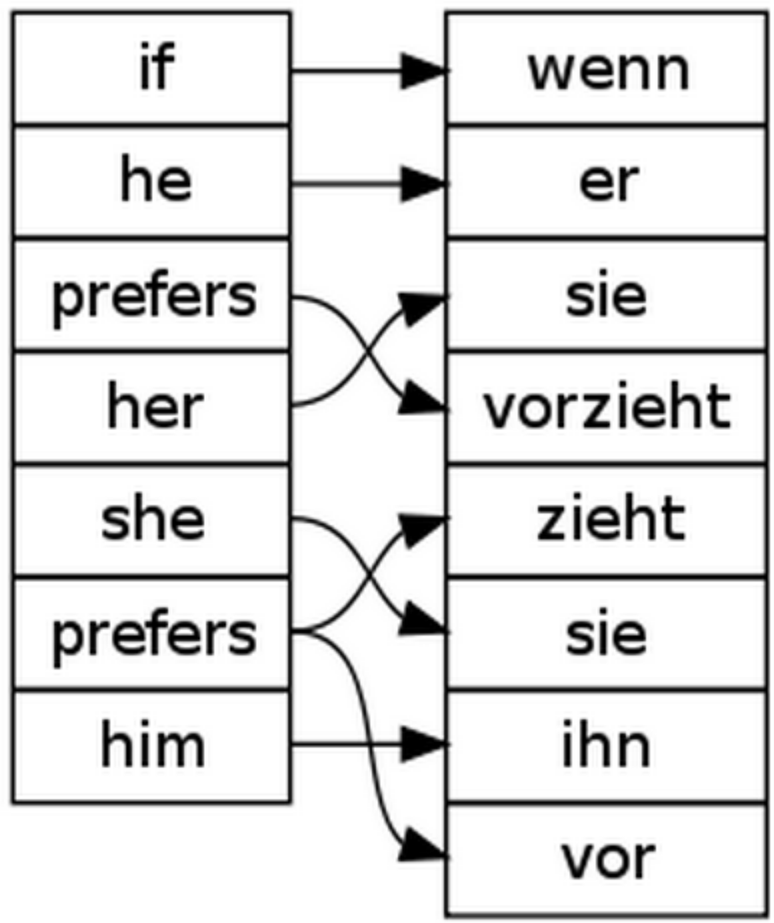
- distances in a tree are logarithmic
- distances are between phrases of arbitrary length
- order is abstract

Data sparsity

- one abstract tree represents many concrete

Phrase alignment via AST

```
If (Pred Prefer He She) (Pred Prefer She He)
```



Where do the trees come from

Abstract syntax trees

- shared between languages
- hiding irrelevant details
  - inflection
  - word order
  - the shape of words
- "pure constituency"
- semantically relevant structure

Cf. word senses in the Universal WordNet

The place of linguistic knowledge

Designing the AST's

Defining the relation between AST and string

- linearization: from trees to strings
- parsing: from strings to trees

The place of statistics

Defining the tree probabilities in each language

This gives us *P(e).*

As trees are shared, we can transfer the probabilities between languages

- in case we don't have data
- justification: "semantics is more or less the same"

What about *P(f|e)*

The probability that the (sought) tree *e* results from the (given) tree *f*.

Intuitively, it has its maximum when *e=f*

But sometimes, this might not do

- *P(e)* might be too low in target language
- the tree may even be missing in the target language, with *P(e)*=0 !

How to find *e* other than *f*

Main rule: *f* and *e* must have the **same meaning**.

We model this as **computational equality** in type theory (a.k.a. **definitional equality**).

Every tree defines a **space of trees** that have the same meaning, via a number of **computation steps**.

Finding the best translation

$$\hat{e} = argmax\ P(f|e)P(e)$$

The best tree *e*, given *f*, maximizes the product of

- target language probability, *P(e)*
- computational closeness to source, *P(f|e)*

under the constraint of sameness of meaning.

Example

Passive transformation elimination Eng-Fin:

Guernica was painted by Picasso

`(Passive Paint Guernica Picasso)`

Guernica maalattiin Picasson toimesta

`(FrontObj Paint Picasso Guernica)`

Guernican maalasi Picasso

What has been done

GF grammar formalism

Resource grammar library: 26 languages

- morphology
- basic syntax
- some language-specific extensions

Probabilistic tree models

- for English from Penn Treebank

First experiments and results

Translating from English to Finnish, German, Hindi, Urdu

Interlingual lexicon based on linked WordNet

(with Shafqat Virk and Krasimir Angelov)

What remains to be done

More coverage for by syntax extensions

Tree models for more languages

Richer notions of tree probability

Complete definition of tree equality

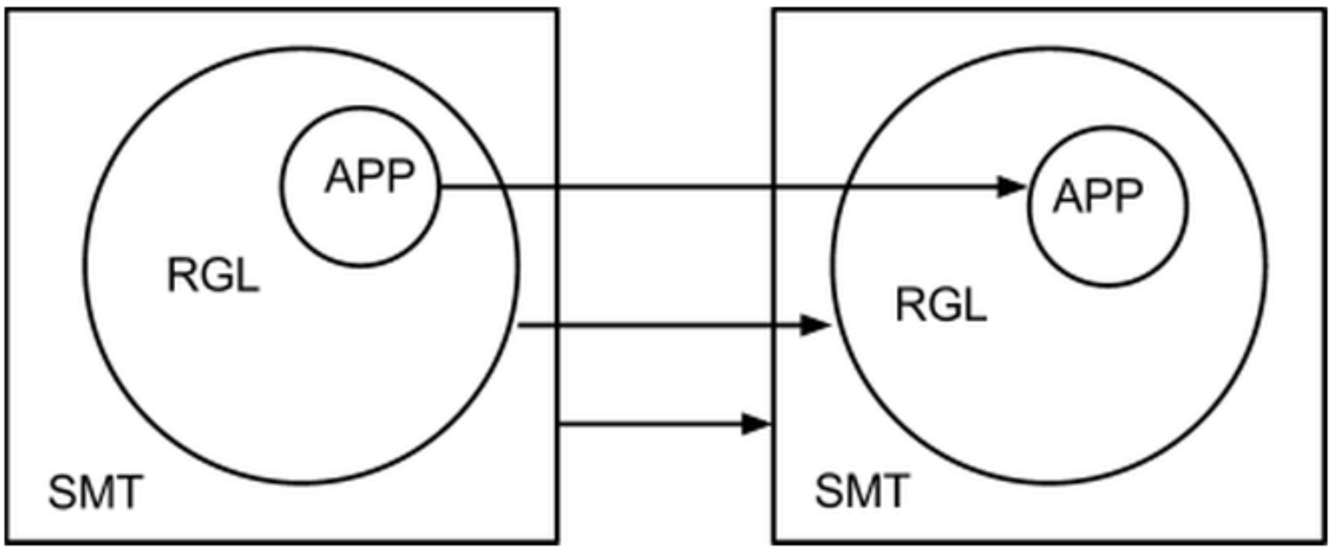
Better back-up models for robustness

Disambiguation with dependent types

Domain and language-pair adaptation

Domain adaptation

Use domain idioms and domain probabilities whenever possible.



Generalization: constructions

how old are you

quelle âge as-tu ("what age do you have")

quanti anni hai ("how many years do you have")

Language-pair adaptation

A short-cut between closely related languages

E.g. the Romance tense system

Or: if a good bilingual lexicon is available

Conclusion

Linguistic knowledge for morphology, syntax, and semantics

Statistical knowledge for idiomacy, robustness, disambiguation

Formalized interlingua, to guarantee

- sameness of meaning
- scalability to many language pairs

Demo

Interlingual translation in GF

http://www.grammaticalframework.org/

Web demo: phrasebook, 19 languages
Web demo: resource grammars, 26 languages
Web demo: robust parsing
Shell demo: hybrid translation, 4 languages
Shell demo: GF resources for Maltese

An experiment

100 fully parsed sentences from Penn Treebank

method/BLEU	Eng-Ger	Eng-Hin	Eng-Fin
Google	**0.63**	0.31	0.29
GF	0.50	**0.34**	**0.35**

Bias: only in-grammar sentences; Penn model

Lecture