Statistical Machine Translation

-introduction-

Artem Sokolov

Computerlinguistik Universität Heidelberg Sommersemester 2015

material from A. Fraser, P. Koehn, S. Vogel, B. Dorr, C. Monz, S. Riezler, C. Botet, H. Thompson

Organization
 Machine Translation

- Vorlesung Artem Sokolov
 - ➡ Thursdays, 11:15-12:45
 - ➡ INF 325 / SR 3
- holiday on Thursdays
 - ➡ 14.05 Himmelfahrt
 - ➡ 04.06 Fronleichnam
- Sprechstunde
 - → Thursdays, 14:00-15:00
 - email beforehand to sokolov@cl...
 - ➡ business trip on 1.06

Übung – Sariya Karimova

- ➡ Tuesdays, 14:15-15:45
- ➡ INF 346 / SR 10
- no sessions after lectures that fall on holidays
 - ➡ 19.05 first Tuesday after Himmelfahrt
 - ➡ 09.06 first Tuesday after Fronleichnam
- Sprechstunde
 - ➡ Wednesdays, 14:00-15:00

- attendance of lectures and practice sessions
- developed SMT system
- homework
- exam 23.07

You will learn:

- basics of learning to translate from corpus data
- basics of internals of mainstream SMT systems
- mathematical details necessary
- analyze the bottlenecks of SMT

1 Softwareprojekt,

Tuesdays, 14:15-17:45 (partial overlap with SMT Übung should be no problem)

 Hauptseminar "Learning and Search in Structured Prediction", Tuesdays, 11:15-12:45

Statistical Machine Translation

Philipp Koehn

questions?

Outline

1 Organization

2 Machine Translation

dreams about automating translation at least since ..th century

- dreams about automating translation at least since ...th century
- some amateur attempts since 1930s

Brief History

- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s
- war anecdotes / 'code talkers':
 - ➡ WW1, 1918, Choctaw
 - ➡ WW2, 1942-1945, Navajo, Basque
 - ➡ Balkans, 1990s, Welsh

- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s
- war anecdotes / 'code talkers':
 - ➡ WW1, 1918, Choctaw
 - ➡ WW2, 1942-1945, Navajo, Basque
 - ➡ Balkans, 1990s, Welsh
- serious projects conceived after 1947
 - ➡ Warren Weaver, "Translation" memorandum
 - language as a code

"This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

- language & invariants (interlingua)
- meaning & context (window context to disambiguate) EN: 'fast' → DE: 'schnell', 'rasch' oder 'bewegungslos', 'fest'
- language & logic
 - (translation as formal "proof" from source "assumptions")
- controlled language (tech. manual, internal docs of corporations)

- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s
- war anecdotes / 'code talkers':
 - ➡ WW1, 1918, Choctaw
 - ➡ WW2, 1942-1945, Navajo, Basque
 - ➡ Balkans, 1990s, Welsh
- serious projects conceived after 1947
 - ➡ Warren Weaver, "Translation" memorandum
 - language as a code

"This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

language & invariants (interlingua)

■ meaning & context (window context to disambiguate) EN: 'fast' → DE: 'schnell', 'rasch' oder 'bewegungslos', 'fest'

language & logic

(translation as formal "proof" from source "assumptions")

controlled language (tech. manual, internal docs of corporations)

first system in 1954 (Georgetown experiment)

- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s
- war anecdotes / 'code talkers':
 - ➡ WW1, 1918, Choctaw
 - ➡ WW2, 1942-1945, Navajo, Basque
 - ➡ Balkans, 1990s, Welsh
- serious projects conceived after 1947
 - ➡ Warren Weaver, "Translation" memorandum
 - language as a code

"This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

language & invariants (interlingua)

■ meaning & context (window context to disambiguate) EN: 'fast' → DE: 'schnell', 'rasch' oder 'bewegungslos', 'fest'

- language & logic (translation as formal "proof" from source "assumptions")
- controlled language (tech. manual, internal docs of corporations)
- first system in 1954 (Georgetown experiment)
- IBM model 1980s

1 commercial

- governments invest in MT languages used by countries that pose economic/military threats
- online translation is VERY popular (the most used of Google's special projects)
- ➡ EU spends more than \$1 billion on translation costs each year
- (semi-)automated translation leads to huge savings for businesses
 - Systran, Unbabel (internships!), Duolingo, Safaba, Fliplingo, ...

1 commercial

- governments invest in MT languages used by countries that pose economic/military threats
- online translation is VERY popular (the most used of Google's special projects)
- ➡ EU spends more than \$1 billion on translation costs each year
- (semi-)automated translation leads to huge savings for businesses
 - Systran, Unbabel (internships!), Duolingo, Safaba, Fliplingo, ...

2 academic

- (probably) the most challenging problem in NLP
- requires knowledge from many NLP sub-areas (semantics, parsing, morphology, stat. modeling)
- enables resource transfer from one language to another over an established link between them



the goal is not to build C-3PO!



gisting

➡ get core message (news digests, hotel reviews)



- gisting and grounding
 - get core message (news digests, hotel reviews)
 - enable action (shopping, booking)



- gisting and grounding
 - get core message (news digests, hotel reviews)
 - enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)

	There is any full labors you can charing with Alassia Vage , your recompersed (all history will appear been,			
The second	Language selection	Chinese (China)		
	Alonso Vega			
Alancalian	Speaks Spanish	Pater .		
 Alonso vega 	writek spanien	Spanish		
available	(9) me			
Translation:	Speaks English			
On 🗾	Writes English			
Speeks Spanish D	e	time		
••••	via Sope Type a message here		0	

Goals of MT

- gisting and grounding
 - get core message (news digests, hotel reviews)
 - enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)
- MT on portable devices (tourists, medical workers, soldiers, augmented reality)



Goals of MT

- gisting and grounding
 - get core message (news digests, hotel reviews)
 - enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)
- MT on portable devices (tourists, medical workers, soldiers, augmented reality)



- gisting and grounding
 - get core message (news digests, hotel reviews)
 - enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)
- MT on portable devices (tourists, medical workers, soldiers, augmented reality)
- support of professional translations
 - rough translation, then post-editing

Des enseignants se rendent régulièrement auprès des élèves de l'institut Jedličkův et leur proposent des activités qui les intéressent et les amusent.

Teachers regularly visit Jedličkův Institute students and offered them activities of interest to them and having fun.

Les étudiants eux-mêmes n'ont pas les moyens de se rendre à des cours, nous essayons de les aider de cette manière.

The students	themselves	cannot be	required	to attend	courses,	we are
trying to hel	themselves	cannot				
	themselves	could not				
Dans le cadre de l' projet dans un no	themselves	do not	instit	'institut Jedlička, nous transférerons ce		
	themselves	cannot af	ford			

- gisting and grounding
 - get core message (news digests, hotel reviews)
 - enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)
- MT on portable devices (tourists, medical workers, soldiers, augmented reality)
- support of professional translations
 - rough translation, then post-editing
 - translation memory

Linguee	■ Dautsch ↔ 🖽 Englisch 🔷 ă ō ū ß				
	raining cats and dogs	Q			
v	Worterbuch Englisch-Deutsch				
Beispiele: It's raining cats and dogs, p_{PI} Siehe auch: rain cats and dogs $v \ll 1$ cats and dogs $p' \ll 1$ ver	(wdw/ 40) → Es regnet in Strömen, (vg Bindlåden regnen v utene spekulätke Wertpapiere βJ (waren/	/ (were) Es regnet Bindfäden, (ky) = c-a © Lingues Wörsebuch, 201			
v	Externe Quell	en (nicht geprüft)			
Do you really want to control cats and dogs, coordinate the tr	the set-up of your event when it is raining schnics, be imitated by capricious [] G+ event-merkening.com	▲ Wollen Sie wirklich den Aufbau bei Wind und Wetter persönlich überwachen, die Technik koordinieren, sich mit kapitidisen (,) Ger ereich warrieting oore			
One day when I walked hom home I saw a little hooded treat	e it was raining cats and dogs. 500 m from ture sitting on the edge (] G+ one-toe-bears.de	Als ich eines Tages heim lief, regnete es wie aus Kübeln. 500 m von Zuhause enfernt sah ich eine kleine Kreatur mit Kapuze, [] Groee-bears.de			
▲ [.] Playa Blanca the sun was shining, while 10 km at the north in Yaza it was raining cats and dogs, lightning and thundering. C+ sviesbe.infe		▲ [.] Situation, dass in Playa Blanca die Sonne schien, während es 10 km nördlich in Yaiza in Kübieln regnete, blizte und donnerte. I2º uzisube örb			
I. J Dubs, Echodub and Tren stations and by DJs around the	chant dubs and gets support on radio globe. It's raining cats and dogs! IDP 2008.elevate.at	▲ [] amerikanischen und britischen Labels ein, ferner das Airplay von Radiostationen und DJs über den gesamten Erdball verteilt. ≅ 2009.einvete ef			
Unfortunately the weather wa dogs, so this picture shows a b	is not on our side - it was raining cats and unch of very wet dogs. G+ milloup.dk	▲ Leider waren die Wettergötter nicht auf unserer Seite, es regnete wie verrückt, so das Bild zeigt eine Gruppe von sehr nassen Hunden.			

- MT task: generate medium- or high-quality translations of documents
- **all** current MT systems work only at sentence level!
- independent translation of sentences is already a very difficult problem
- important discourse phenomena are ignored: Example: How to translate English 'it' to German (feminine/masculine/neutral) if object referred to was in previous sentences?

Approaches to MT

- grammar-based / rule-based
 - ➡ interlingua
 - transfer
- direct
 - statistical
 - example-based



Approaches to MT

- grammar-based / rule-based
 - ➡ interlingua
 - transfer
- direct
 - statistical
 - example-based



Approaches to MT

- grammar-based / rule-based
 - ➡ interlingua
 - transfer
- direct
 - statistical
 - example-based



- using statistical models
 - create many alternatives, hypotheses
 - ➡ give a score to each hypothesis
 - ightarrow select the best ightarrow search

- using statistical models
 - create many alternatives, hypotheses
 - give a score to each hypothesis
 - ightarrow select the best ightarrow search
- advantages
 - avoids hard decisions
 - speed can be traded with quality, no all-or-nothing
 - works better in the presence of unexpected/disfluent input
 - ➡ learns from real world, abundant data
 - high model and methods reusability

- using statistical models
 - create many alternatives, hypotheses
 - ➡ give a score to each hypothesis
 - ightarrow select the best ightarrow search
- advantages
 - avoids hard decisions
 - speed can be traded with quality, no all-or-nothing
 - works better in the presence of unexpected/disfluent input
 - ➡ learns from real world, abundant data
 - high model and methods reusability
- disadvantages
 - difficulties handling structurally rich models, mathematically and computationally
 - need more data to train the model with increasing number of parameters
 - ➡ not easily interpretable, difficult to distill rules by observing the system

How to Build an SMT System

Training:

1 large parallel corpus

➡ consists of document pairs (document and its translation)

Training:

1 large parallel corpus

- consists of document pairs (document and its translation)
- **2 sentence alignment**: in each document pair find those sentences which are translations of one another
 - ➡ results in sentence pairs (sentence and its translation)

Training:

1 large parallel corpus

- consists of document pairs (document and its translation)
- **2 sentence alignment**: in each document pair find those sentences which are translations of one another
 - results in sentence pairs (sentence and its translation)
- **3 word alignment**: in each sentence pair annotate those words which are translations of one another
 - results in aligned word-phrases
1 large parallel corpus

- consists of document pairs (document and its translation)
- **2 sentence alignment**: in each document pair find those sentences which are translations of one another
 - results in sentence pairs (sentence and its translation)
- **3 word alignment**: in each sentence pair annotate those words which are translations of one another
 - results in aligned word-phrases
- 4 estimate a **statistical model** from the word-aligned sentence pairs
 - results in translation model parameters

1 large parallel corpus

- consists of document pairs (document and its translation)
- **2 sentence alignment**: in each document pair find those sentences which are translations of one another
 - results in sentence pairs (sentence and its translation)
- **3 word alignment**: in each sentence pair annotate those words which are translations of one another
 - results in aligned word-phrases
- 4 estimate a statistical model from the word-aligned sentence pairs
 - results in translation model parameters

Language Modeling:

1 large parallel corpus

- consists of document pairs (document and its translation)
- **2 sentence alignment**: in each document pair find those sentences which are translations of one another
 - results in sentence pairs (sentence and its translation)
- **3 word alignment**: in each sentence pair annotate those words which are translations of one another
 - results in aligned word-phrases
- 4 estimate a statistical model from the word-aligned sentence pairs
 - results in translation model parameters

Language Modeling:

- 5 large monolingual corpus
 - ➡ texts in target language

1 large parallel corpus

- consists of document pairs (document and its translation)
- **2 sentence alignment**: in each document pair find those sentences which are translations of one another
 - results in sentence pairs (sentence and its translation)
- **3 word alignment**: in each sentence pair annotate those words which are translations of one another
 - results in aligned word-phrases
- 4 estimate a statistical model from the word-aligned sentence pairs
 - results in translation model parameters

Language Modeling:

- 5 large monolingual corpus
 - ➡ texts in target language
- **6** estimate a **statistical model** from examples of well-formed language
 - results in language model: how likely a word will follow a given history

Tuning:

6 define how important is every model for translation quality

➡ results in a complete model

Tuning:

6 define how important is every model for translation quality

➡ results in a complete model

Testing:

given new text to translate, apply model to get most likely translation

Tuning:

6 define how important is every model for translation quality

➡ results in a complete model

Testing:

given new text to translate, apply model to get most likely translation



Traditional focus was on high-resourced languages:

- high demand ⇒ data collection efforts
- available data \Rightarrow spawns research
- quality systems \Rightarrow proliferation, new markets \Rightarrow more demand

Traditional focus was on high-resourced languages:

- high demand ⇒ data collection efforts
- available data \Rightarrow spawns research
- quality systems \Rightarrow proliferation, new markets \Rightarrow more demand

No clear-cut definition in number of words:

$> 200 \mathrm{M}$ high-resourced	French, Chinese, Arabic
$\sim 50 {\rm M}$ medium-resourced	German, Portuguese, Italian
$< 5 {\rm M}$ under-resourced	Tatar, Uzbek, Estonian
$< 100 {\rm K}$ close to none	Chechen, Udmurt, <i>Silbo</i> , <i>Klingon</i> :)
heavily depends on a language pa	ir and direction:

for example: ZH-EN is well-resourced, FR-ZH is much less so

english	german
Diverging opinions about planned	Unterschiedliche Meinungen zur
tax reform	geplanten Steuerreform
The discussion around the envis-	Die Diskussion um die vorgesehene
aged major tax reform continues .	grosse Steuerreform dauert an .
The FDP economics expert, Graf	Der FDP - Wirtschaftsexperte
Lambsdorff , today came out in fa-	Graf Lambsdorff sprach sich heute
vor of advancing the enactment of	dafuer aus , wesentliche Teile
significant parts of the overhaul ,	der fuer 1999 geplanten Reform
currently planned for 1999 .	vorzuziehen .

english	german
Diverging opinions about planned	Unterschiedliche Meinungen zur
tax reform	geplanten Steuerreform
The discussion around the envis-	Die Diskussion um die vorgesehene
aged major tax reform continues .	grosse Steuerreform dauert an .
The FDP economics expert , Graf	Der FDP - Wirtschaftsexperte
Lambsdorff , today came out in fa-	Graf Lambsdorff sprach sich heute
vor of advancing the enactment of	dafuer aus , wesentliche Teile
significant parts of the overhaul , currently planned for 1999 .	der fuer 1999 geplanten Reform vorzuziehen .

note some pre-processing (tokenization, normalization)

- if document D_e is translation of document D_f , how to find the translation for each sentence?
- the *n*-th sentence in D_e is not necessarily the translation of the *n*-th sentence in D_f
- in addition to 1:1 alignments, there are also 1:0, 0:1, 1:n, and n:1
- in EuroParl proceedings, $\sim 90\%$ of the sentence alignments are 1:1



given sentences that are translation of one another, how to know which words are mutual translations?





given sentences that are translation of one another, how to know which words are mutual translations?



Goal:

- $\blacksquare\$ get a score function p(e|f) goodness of translation e given foreign input f
 - 1 p('die Waschmaschine läuft', 'the washing machine is running') = 0.952 p('die Waschmaschine läuft', 'the car drove') = 0.03
- convenient to think of p as probability
- models to some extent natural language's uncertainty and ambiguity
- translation: $\arg \max_{e} p(e|f)$

What kind of function can p(e|f) be?:

- one naïve way to determine p(e|f):
 - **1** count how many times f was translated by e_1 or e_2 in the training data

2 set
$$p(e_1|f) = \frac{\#\{f \to e_1\}}{\#\{f \to ?\}}$$

3 set
$$p(e_2|f) = \frac{\#\{f \to e_2\}}{\#\{f \to ?\}}$$

- only works of we saw exactly the f and e_1, e_2 in our training data
- we can't generalize to unseen sentences
- solution decompose input and output into parts

4. Translation Model - Maximum Likelihood Estimation



- generate a word alignment for each sentence pair
- count the number of times every source word was linked to every target word:

1
$$\#$$
{das \rightarrow the} = 1

2
$$\#$$
{Haus \rightarrow house} = 1

- $3 \#{ist \rightarrow is} = 1$
- 4 #{klitzeklein \rightarrow very} = 1
- **5** #{klitzeklein \rightarrow small} = 1

4. Translation Model - Maximum Likelihood Estimation



- generate a word alignment for each sentence pair
- count the number of times every source word was linked to every target word:

2
$$\#$$
{Haus \rightarrow house} = 1.0

- **3** $\#{\text{ist} \to \text{is}} = 1.0$
- 4 #{klitzeklein \rightarrow very} = 0.5
- 5 #{klitzeklein \rightarrow small} = 0.5
- divide by the number of occurrences of the source word

4. Translation Model - Maximum Likelihood Estimation



- generate a word alignment for each sentence pair
- count the number of times every source word was linked to every target word:

1
$$#{das \rightarrow the} =$$

- 2 #{Haus \rightarrow house} =
- $3 \#{ist \rightarrow is} =$
- 4 #{klitzeklein \rightarrow very} =
- 5 #{klitzeklein \rightarrow small} =
- divide by the number of occurrences of the source word
- this is our word/phrase translation probability $p(w_e|w_f)$

- decomposing can introduce output disfluencies
- need to somehow improve fluency in translations
- learn what is "fluent" from examples of well-formed language
- results in language model: how likely a word will follow a given history
 - ➡ p(Haus|Das kleine) > p(Haus|Die kleine)

Translating is usually referred to as decoding (W. Weaver, 1947)



SMT was born from automatic speech recognition:

- p(e) =language model
- p(f|e) = acoustic model

Translating is usually referred to as decoding (W. Weaver, 1947)



SMT was born from automatic speech recognition:

- p(e) =language model
- p(f|e) = acoustic model

Decoding

Translating is usually referred to as decoding (W. Weaver, 1947)



SMT was born from automatic speech recognition:

- $\ \ \, {\bf P}(e) = {\rm language \ model}$
- p(f|e) = acoustic model
- however, SMT must deal with word reordering!

$$\arg\max_{e} p(e|f) = \arg\max_{e} p(f|e)p(e)$$

$$\underset{e}{\arg\max \log p(e|f)} = \underset{e}{\arg\max \log p(f|e)} + \log p(e)$$

move to log-space

$$\underset{e}{\arg\max \log p(e|f)} = \underset{e}{\arg\max \alpha \log p(f|e)} + \beta \log p(e)$$

- move to log-space
- models may have different importance (weight)

$$\underset{e}{\arg\max} \log p(e|f) = \underset{e}{\arg\max} \alpha \log p(f|e) + \beta \log p(e) + \gamma f(\cdot)$$

- move to log-space
- models may have different importance (weight)
- we may want to add more models

$$\underset{e}{\arg\max \log p(e|f)} = \underset{e}{\arg\max \alpha f_1(\cdot)} + \beta f_2(\cdot) + \gamma f_3(\cdot)$$

- move to log-space
- models may have different importance (weight)
- we may want to add more models
- they even need not to be log-probabilities (features)

Generalization

$$\arg\max_{e} \log p(e|f) = \arg\max\sum_{i=1}^{n} w_i f_i(e, f)$$

- move to log-space
- models may have different importance (weight)
- we may want to add more models
- they even need not to be log-probabilities (features)
- maximize score function a weighted linear combination of features



- **1** find such w_i that maximize translation quality
- 2 many methods exist and still an active research area

- how to know if your SMT system works well?
- run it on a large number of unseen sentences and evaluate the quality
- but what is 'quality'?
 - ⇒ can evaluate MT at corpus, document, sentence or word level..
 - ➡ in the MT the unit of translation is the sentence
- human evaluation of MT quality is difficult (expensive)
- need an abstract measure of usefulness of the output
 - evaluation metric: assigns a score to a hypothesized translation
 - automatic evaluation metrics rely on comparison with selected human translations

WER (word error rate)

- ➡ edit distance to reference translation (insertion, deletion, substitution)
- ➡ captures fluency well, adequacy not so well
- ➡ rigid: gives no credit for translating 'Frau' instead of 'Fraulein'
- **TER** (translation error rate)
 - edit distance to reference translation (+ block moves)
 - ➡ captures reordering freedom better, very good correlation with humans
 - ➡ common problems: synonyms,
- BLEU (most popular)
 - ➡ counts matching n-grams
 - ➡ captures fluency, rewards long and fluent matches
 - penalizes the noisy channel model's tendency to produce short outputs
 - ➡ well-correlates with humans, very intuitive, easier then TER for learning
 - cons: no credit for synonyms, for legitimate but slightly reordered outputs

METEOR

- combines synonyms, stemming, WordNet synsets
- ➡ most "human like"
- attempts to capture language flexibility
- cons: language dependent (stemmer, WordNet)

Our groups research directions cross-lingual information retrieval (e.g., patents)

- **1** cross-lingual information retrieval (e.g., patents)
- 2 grounded learning

- **1** cross-lingual information retrieval (e.g., patents)
- 2 grounded learning
- 3 learning from weak feedback ('under-paid turkers')

- **1** cross-lingual information retrieval (e.g., patents)
- 2 grounded learning
- 3 learning from weak feedback ('under-paid turkers')
- 4 learning in non-cooperative environment

- **1** cross-lingual information retrieval (e.g., patents)
- 2 grounded learning
- 3 learning from weak feedback ('under-paid turkers')
- 4 learning in non-cooperative environment
- 5 learning from non-parallel data
Our groups research directions

- **1** cross-lingual information retrieval (e.g., patents)
- 2 grounded learning
- 3 learning from weak feedback ('under-paid turkers')
- 4 learning in non-cooperative environment
- 5 learning from non-parallel data
- 6 SMT with neural networks

Our groups research directions

- **1** cross-lingual information retrieval (e.g., patents)
- 2 grounded learning
- 3 learning from weak feedback ('under-paid turkers')
- 4 learning in non-cooperative environment
- 5 learning from non-parallel data
- 6 SMT with neural networks
- include over-sentential context

Challenges

Our groups research directions

- 1 cross-lingual information retrieval (e.g., patents)
- 2 grounded learning
- 3 learning from weak feedback ('under-paid turkers')
- 4 learning in non-cooperative environment
- 5 learning from non-parallel data
- 6 SMT with neural networks
- include over-sentential context

SMT projects from this term's SWP (today, 16:15, INF327 SR2):

- quasi-parallel corpus creation
- kernel-SMT without alignments
- SMT on character levels
- neural networks for bilingual word representations
- user feedback based SMT learning

- 1 Word-Based Models
- 2 Phrase-Based SMT
- 3 Decoding
- 4 Language Models
- 5 Evaluation
- 6 Tree-Based SMT

see you the day after tomorrow at 11:15, INF 327 SR3