Introduction to Statistical Machine Translation

ESSLLI 2005

Chris Callison-Burch
Philipp Koehn

Who we are

- Chris = PhD student at University of Edinburgh, co-founder of Linear B Ltd, a startup company that builds SMT systems
- Philipp = Lecturer at U of Edinburgh, recently finished his PhD at University of Southern California / ISI, did postdoc at MIT

Course Overview

- Day 1:
  - Different approaches to MT
  - Overview of statistical MT
  - Useful resources
- Day 2:
  - Decoding and search
- Day 3:
  - Aligning words and phrases

Course Overview

- Day 4:
  - Evaluation of translation quality
  - Using parallel corpora for other tasks
- Day 5:
  - Syntax-based approaches to SMT

Overview of MT

A long history

- Machine translation was one of the first applications envisioned for computers
- Warren Weaver (1949)
  "I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text."
- First demonstrated by IBM in 1954 with a basic word-for-word translation system.

Commercially Interesting

- U.S. has invested in MT for intelligence purposes
- MT is popular on the web -- it is the most used of Google's special features
- EU spends more than €1,000,000,000 on translation costs each year. (Semi-) automating that could lead to huge savings

Academically Interesting

- Machine translation requires many other NLP technologies
- Potentially: parsing, generation, word sense disambiguation, named entity recognition, transliteration, pronoun resolution, natural language understanding, and real-world knowledge

What makes MT hard?

- Word order
- Word sense
- Pronouns
- Tense
- Idioms

Differing word orders

- English word order is *subject - verb - object*
  Japanese order is *subject - object - verb*
- English: IBM bought Lotus
  Japanese: *IBM Lotus bought*
- English: Reporters said IBM bought Lotus
  Japanese: *Reporters IBM Lotus bought said*

Word sense ambiguity

- `Bank' as in river
  `Bank' as in financial institution
- `Plant' as in a tree
  `Plant' as in a factory
- Different word senses will likely translate into different words in another language

Problem of pronouns

- Some languages like Spanish can drop subject pronouns
- In Spanish the verbal inflection often indicates which pronoun should be restored
  -o = I
  -as = you
  -a = he / she / it
  -amos = we
  -an = they
- When should we use `she' or `he' or `it'?

Different tenses

- Spanish has two versions of the past tense: one for a definite time in the past, and one for an unknown time in the past
- When translating from English to Spanish we need to choose which version of the past tense to use

Idioms

- "to kick the bucket" means "to die"
- "a bone of contention" does not have anything to do with skeletons
- "a lame duck", "tongue in cheek", "to cave in"

Various Approaches to Machine Translation

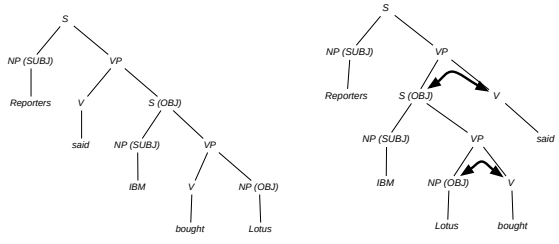
Various approaches

- Word-for-word translation
- Syntactic transfer
- Interlingual approaches
- Controlled language
- Example-based translation
- Statistical translation

Word-for-word translation

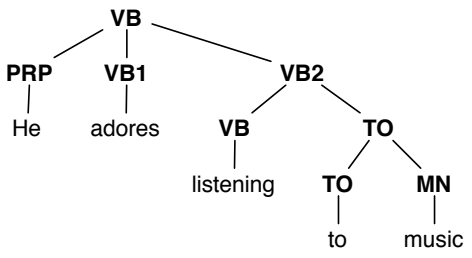
- Use a machine-readable bilingual dictionary to translate each work in a text
- Advantages: Easy to implement, results give a rough idea about what the text is about
- Disadvantages: Problems with word order means that this results in low-quality translation

Syntactic transfer

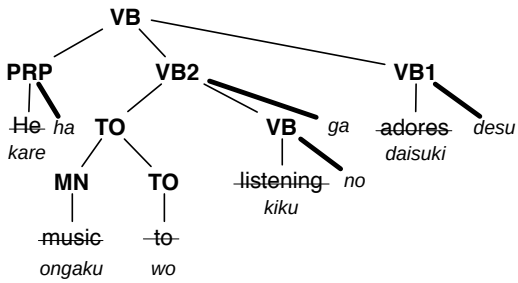


- Parse the sentence
- Rearrange constituents
- Then translate the words

Syntactic transfer



Syntactic transfer



Syntactic transfer

- Advantages:
  Deals with the word-order problem
- Disadvantages:
  - Must construct transfer rules for each language pair that you deal with
  - Sometimes there is syntactic mis-match
- Example:
  English: The bottle floated into the cave
  Spanish: *La botella entro a la cuerva flotando*
  = *The bottle entered the cave floating*

Interlingua

- Assign a logical form to sentences
- John must not go =
  `OBLIGATORY(NOT(GO(JOHN)))`
  John may not go =
  `NOT(PERMITTED(GO(JOHN)))`
- Use logical form to generate a sentence in another language

Interlingua

- Advantages:
  Single logical form means that we can translate between all languages and only write a parser/generator for each language once
- Disadvantages:
  Difficult to define a single logical form. English words in all capital letter probably won't cut it.

Controlled language

- Define a subset of a language which can be used to compose text to be translated
- Issued editorial guidelines that limit each word to only one word sense, and which forbid certain difficult constructions
- Apply syntactic transfer or interlingual approaches

Controlled language

- Advantages: Results in more reliable, higher quality translation for subset of language that it deals with
- Disadvantages: Does not cover all language use, so can only be applied in limited settings

Example-based MT

- Fundamental idea:
  - People do not translate by doing deep linguistics analysis of a sentence.
  - They translate by decomposing sentence into fragments, translating each of those, and then composing those properly.
- Principle of analogy in translation

Example of Example-Based MT

- Translate:
  He buys a book on international politics.
- With these examples:
  (He buys) a notebook.
  (Kare ha) nouto (wo kau).

  I read (a book on international politics).
  Watashi ha (kokusaiseiji nitsuite kakareta hon) wo yomu
- *(Kare ha) (kokusaiseiji nitsuite kakareta hon) (wo kau).*

Challenges

- Locating similar sentences
- Aligning sub-sentential fragments
- Combining multiple fragments of example translations into a single sentence
- Determining when it is appropriate to substitute one fragment for another
- Selecting the best translation out of many candidates

Example-based MT

- Advantages: Uses fragments of human translations which can result in higher quality
- Disadvantages: May have limited coverage depending on the size of the example database, and flexibility of matching heuristics

Statistical machine translation

- Find most probable English sentence given a foreign language sentence
- Automatically align words and phrases within sentence pairs in a parallel corpus
- Probabilities are determined automatically by training a statistical model using the parallel corpus

Parallel corpus

what is more , the relevant cost dynamic is completely under control.	im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle .
sooner or later we will have to be sufficiently progressive in terms of own resources as a basis for this fair tax system .	früher oder später müssen wir die notwendige progressivität der eigenmittel als grundlage dieses gerechten steuersystems zur sprache bringen .
we plan to submit the first accession partnership in the autumn of this year .	wir planen , die erste beitrittspartnerschaft im herbst dieses jahres vorzulegen .
it is a question of equality and solidarity .	hier geht es um gleichberechtigung und solidarität .
the recommendation for the year 1999 has been formulated at a time of favourable developments and optimistic prospects for the european economy .	die empfehlung für das jahr 1999 wurde vor dem hintergrund günstiger entwicklungen und einer für den kurs der europäischen wirtschaft positiven perspektive abgegeben .
that does not , however , detract from the deep appreciation which we have for this report .	im übrigen tut das unserer hohen wertschätzung für den vorliegenden bericht keinen abbruch .

Statistical machine translation

- Advantages:
  - Has a way of dealing with lexical ambiguity
  - Can deal with idioms that occur in the training data
  - Requires minimal human effort
  - Can be created for any language pair that has enough training data
- Disadvantages:
  Does not explicitly deal with syntax

Choosing an Approach

- Many challenges in MT, many different ways of approaching the task
- What approach you prefer will depend on your background (i.e. logicians tend towards interlingua, linguists towards syntactic transfer)
- Objectively choosing how to approach the task is tricky

Some Criteria

- Do we want to design a system for a single language or for many languages?
- Can we assume a constrained vocabulary or do we need to deal with any text?
- What resources already exist for the languages that we're dealing with?
- How long will it take us to develop the resources, and how large a staff will we need?

Advantages of SMT

- Data driven
- Language independent
- No need for staff of linguists of language experts
- Can prototype a new system quickly and at a very low cost

Choosing SMT

- Economic reasons:
  - Low cost; Rapid prototyping
- Practical reasons:
  - Many language pairs don't have NLP resources, but do have parallel corpora
- Quality reasons:
  - Uses chunks of human translated as its building blocks
  - When very large data sets are available produces state of the art results

Overview of SMT

Probabilities

- Find most probable English sentence given a foreign language sentence

$$p(e|f)$$

$$\hat{e} = \arg\max_e \; p(e|f)$$

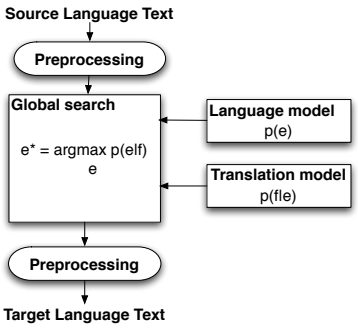
$$p(e|f) = \frac{p(e)p(f|e)}{p(f)}$$

$$\hat{e} = \arg\max_e \; p(e)p(f|e)$$

What the probabilities represent

- p(e) is the "Language model"
  - Assigns a higher probability to fluent / grammatical sentences
  - Estimated using monolingual corpora
- p(f|e) is the "Translation model"
  - Assigns higher probability to sentences that have corresponding meaning
  - Estimated using bilingual corpora

For people who don't like equations



Language Model

- Component that tries to ensure that words come in the right order
- Some notion of grammaticality
- Standardly calculated with a trigram language model, as in speech recognition
- Could be calculated with a statistical grammar such as a PCFG

Trigram language model

- p(I like bungee jumping off high bridges) =
  p(I | <s> <s>) *
  p(like | I <s>) *
  p(bungee | I like) *
  p(jumping | like bungee) *
  p(off | bungee jumping) *
  p(high | jumping off) *
  p(bridges | off high) *
  p(</s> | high bridges) *
  p(</s> | bridges </s>)

Calculating Language Model Probabilities

- Unigram probabilities

$$p(w_1) = \frac{count(w_1)}{total\ words\ observed}$$

Calculating Language Model Probabilities

- Bigram probabilities

$$p(w_2|w_1) = \frac{count(w_1 w_2)}{count(w_1)}$$

Calculating Language Model Probabilities

- Trigram probabilities

$$p(w_3|w_1 w_2) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)}$$

Calculating Language Model Probabilities

- Can take this to increasingly long sequences of n-grams
- As we get longer sequences it's less likely that we'll have ever observed them

Backing off

- Sparse counts are a big problem
- If we haven't observed a sequence of words then the count = 0
- Because we're multiplying the n-gram probabilities to get the probability of a sentence the whole probability = 0

Backing off

- $.8 * p(w_3|w_1 w_2) +$
  $.15 * p(w_3|w_2) +$
  $.049 * p(w_3) +$
  $.001$
- Avoids zero probs

Translation model

- p(f|e)... the probability of some foreign language string given a hypothesis English translation
- f = Ces gens ont grandi, vécu et oeuvré des dizaines d'années dans le domaine agricole.
- e = *Those people have grown up, lived and worked many years in a farming district.*
- e = *I like bungee jumping off high bridges.*

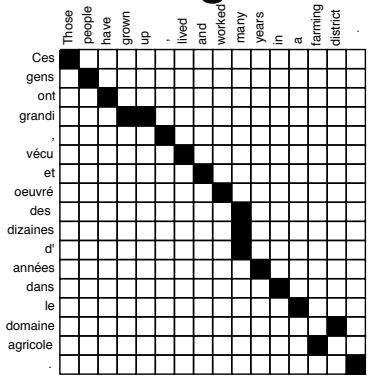
Translation model

- How do we assign values to p(f|e)?
- $$p(f|e) = \frac{count(f, e)}{count(e)}$$
- Impossible because sentences are novel, so we'd never have enough data to find values for all sentences.

Translation model

- Decompose the sentences into smaller chunks, like in language modeling
- $$p(f|e) = \sum_a p(a, f|e)$$
- Introduce another vairable *a* that represents alignments between the individual words in the sentence pair

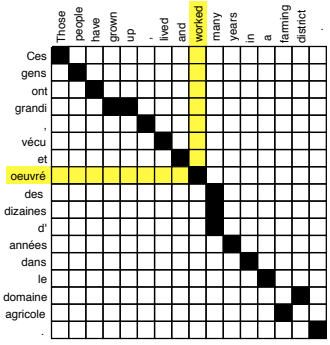
Word alignment



Alignment probabilities

- So we can calculate translation probabilities by way of these alignment probabilities
$$p(f|e) = \sum_a p(a, f|e)$$
- Now we need to define p(a, f | e)
$$p(a, f|e) = \prod_{j=1}^{m} t(f_j|e_i)$$

Calculating $t(f_j|e_i)$

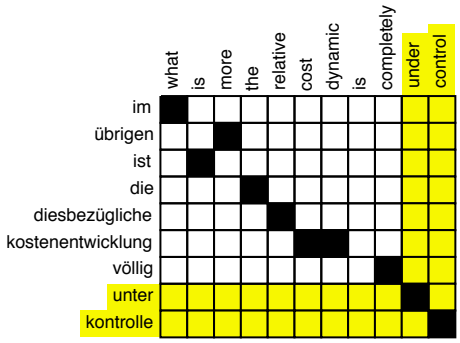


- Counting! I told you probabilities were easy!
- $= \frac{count(f_j, e_i)}{count(e_i)}$
- **worked... fonctionné, travaillé, marché, oeuvré**
- 100 times total 13 with this f. 13%

Calculating $t(f_j|e_i)$

- Unfortunately we don't have word aligned data, so we can't do this directly.
- OK, so it's not quite as easy as I said.
- Philipp will talk about how to do word alignments using EM on Wednesday.

Phrase Translation Probabilities



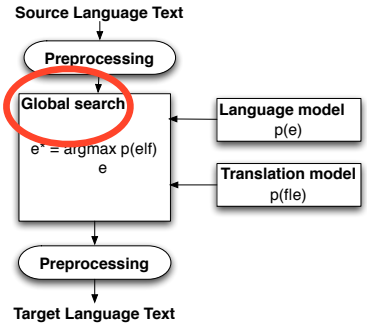
Phrase Translation Probabilities

Phrase Table

- Exhaustive table of source language phrases paired with their possible translations into the target language, along with probabilities

das thema	the issue	.51
	the point	.38
	the subject	.21

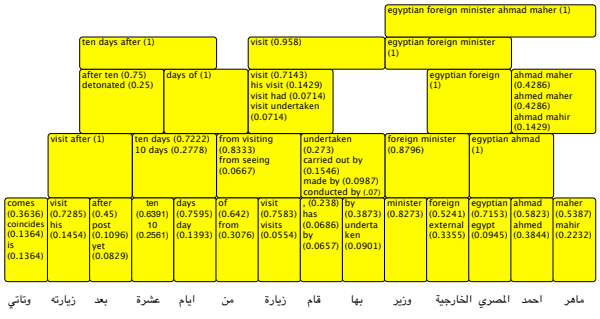
``Diagram Number 1"



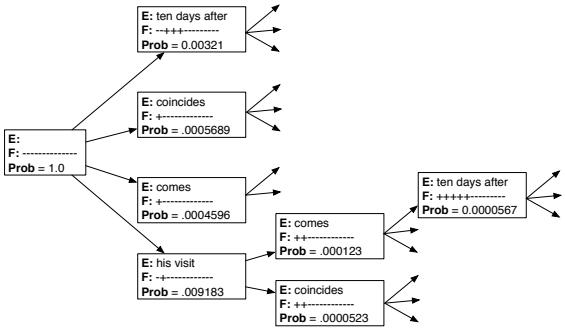
The Search Process AKA ``Decoding"

- Look up all translations of every source phrase, using the phrase table
- Recombine the target language phrases that maximizes the translation model probability * the language model probability
- This search over all possible combinations can get very large so we need to find ways of limiting the search space

Looking up translations of source



The Search Space



The Search Space

- In the end the item which covers all of the source words and which has the highest probability wins!
- That's our best translation

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

- And there was much rejoicing!

Wrap-up: SMT is data driven

- Learns translations of words and phrases from parallel corpora
- Associate probabilities with translations empirically by counting co-occurrences in the data
- Estimates of probabilities get more accurate as size of the data increases

Wrap-up: SMT is language independent

- Can be applied to any language pairs that we have a parallel corpus for
- The only linguistic thing that we need to know is how to split into sentences, words
- Don't need linguists and language experts to hand craft rules because it's all derived from the data

Wrap-up: SMT is cheap and quick to produce

- Low overhead since we aren't employing anyone
- Computers do all the heavy lifting / statistical analysis of the data for us
- Can build a system in around 2 weeks

Example translations

Spanish --> English

- Sabemos muy bien que los tratados actuales no bastan y que, en el futuro, será necesario desarrollar una estructura mejor y diferente para la unión europea, una estructura más constitucional que también deje bien claras cuáles son las competencias de los estados miembros y cuáles pertenecen a la unión.
- We all know very well that the current treaties are insufficient and that, in the future, it will be necessary to develop a better structure and different for the European Union, a structure more constitutional also make it clear what the competences of the member states and which belong to the union.

German --> English

- Uns ist sehr wohl bewusst, dass die geltenden verträge unzulänglich sind und künftig eine andere, effizientere struktur für die union entwickelt werden muss, nämlich eine stärker konstitutionell ausgeprägte struktur mit einer klaren abgrenzung zwischen den befugnissen der mitgliedstaaten und den kompetenzen der union.
- We are well aware that the existing treaties are inadequate and in the future, a different, more efficient structure for the union must be developed, namely a more pronounced institutional structure with a clear dividing line between the powers of the member states and the competences of the union.

Danish --> English

- Vi ved udmærket, at de nuværende traktater ikke er tilstrækkelige, og at det i fremtiden er nødvendigt at udvikle en anden og bedre struktur for unionen, en mere konstitutionel struktur, som også tydeligt viser, hvilke beføjelser medlemsstaterne har, og hvilke beføjelser unionen har.
- We know perfectly well that the current treaties are not sufficient, and that in the future it is necessary to develop a second and better structure for the union, a more constitutional structure, which clearly shows the powers of the member states, and what powers the union.

French --> English

- Nous savons très bien que les traités actuels ne suffisent pas et qu'il sera nécessaire à l'avenir de développer une structure plus efficace et différente pour l'union, une structure plus constitutionnelle qui indique clairement quelles sont les compétences des états membres et quelles sont les compétences de l'union.
- We know very well that the current treaties are not enough and that it will be necessary in future to develop a structure more effective and different for the union, a more constitutional structure which makes it clear what are the competence of member states and what are the powers of the union.

Spanish --> English (2)

- Mensajes de preocupación en primer lugar ante las dificultades económicas y sociales por las que atravesamos, y ello a pesar de un crecimiento sostenido, fruto de años de esfuerzo por parte de todos nuestros conciudadanos.
- Messages of concern in the first place just before the economic and social problems for the present situation, and in spite of sustained growth, as a result of years of effort on the part of our citizens.

German --> English (2)

- Dabei handelt es sich zunächst um botschaften der beunruhigung angesichts der wirtschaftlichen und sozialen schwierigkeiten, mit denen wir trotz eines anhaltenden wachstums als ergebnis jahrelanger anstrengungen von seiten aller unserer mitbürger konfrontiert sind.
- It is, first of all, embassies of the concern about the economic and social problems, with which we, despite the continuing growth as a result of many years of effort from all of our fellow citizens are confronted.

Danish --> English (2)

- Vi må videregive et budskab om bekymring set i lyset af de økonomiske og sociale problemer, vi aktuelt oplever, uanset at der meldes om stabil økonomisk vækst, hvilket må ses som resultatet af den indsats, der de seneste år er ydet af eu's borgere.
- We must convey a message of concern in the light of the economic and social problems, we topical experiencing, regardless of the fact that there are reported on stable economic growth, which must be seen as the result of the efforts that the last few years done by the EU's citizens.

French --> English (2)

- Messages d'inquiétude tout d'abord devant les difficultés économiques et sociales que nous traversons, et ce malgré une croissance soutenue, fruit d'années d'efforts de la part de tous nos concitoyens.
- Messages of concern firstly to the economic and social problems that we are going through, despite sustained growth as a result of years of effort on the part of all our citizens.

English References

- We know all too well that the present treaties are inadequate and that the union will need a better and different structure in future, a more constitutional structure which clearly distinguishes the powers of the member states and those of the union.
- These are, first and foremost, messages of concern at the economic and social problems that we are experiencing, in spite of a period of sustained growth stemming from years of efforts by all our fellow citizens.

Useful Resources

Materials Needed to Build an SMT System

- Parallel corpus
- Word alignment software
- Language modeling toolkit
- Decoder

Parallel Corpora

- The Linguistics Data consortium sells many parallel corpora including
  - UN data
  - Canadian Hansards
  - Hong Kong laws parallel text
  - Parallel newswires
- http://www.ldc.upenn.edu/

Parallel Corpora

- Philipp has the "Europarl Corpus"
  - Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish
- http://www.iccs.informatics.ed.ac.uk/~pkoehn/publications/europarl/

Word alignment software

- Giza++ (open source implementation of the "IBM Models")
- http://www.fjoch.com/GIZA++.html
- Reference word alignments
  - Manually created "gold standard"
  - Useful for testing the quality of automatically generated alignments
  - See ACL-05 / NAACL-03 workshops

Language modeling toolkit

- SRILM
  - Developed for speech recognition
  - Used in SMT too
  - Estimates n-gram probabilities
  - Handles back-off in sophisticated ways
- http://www.speech.sri.com/projects/srilm/

Decoder

- Pharaoh
  - Phrase-based SMT decoder
  - Builds phrase tables from Giza++ word alignments
  - Produces best translation for new input using phrase table plus SRILM language model
- http://www.isi.edu/licensed-sw/pharaoh/

Tomorrow

- Decoding in depth
- Including: How to use Pharaoh!