CMPT 825: Natural Language Processing **Spring 2008**

Lecture 9 — March 3-8, 2008

Lecturer: Anoop Sarkar *Scribe: Baskaran Sankaran*

Lectures Outline:

1. Two aspects: Fluency and content transfer
   - Language model
   - Translation model
2. Translation process
   - Fertility
   - Null-insertion probability
   - Translation
   - Distortion (reordering)
3. Evolution of SMT
   - Word-based models
   - Phrase-based models
   - Syntax-based models
4. Evaluation of MT systems
   - Automatic evaluation
     - WER (from Speech processing community)
     - BLEU
     - Correlation of BLEU with human evaluation
   - Experiments by Philip Koehn
     - Disadvantages of BLEU (going out of vs. going in to English)
5. Sidewalk
   - MT pyramid
   - Rule-based approaches
6. Translation modeling
   - Noisy-channel model
   - IBM Model 3

9.1 A simplified view of MT

Consider the Portuguese sentence given below and five different English *translations* for it.
Portuguese: Atentado suicida mata 5 soldados dos EUA em Bagd

English-1 : Five US soldiers killed in Baghdad
English-2 : 5 US soldiers killed in suicide bombing in Baghdad
English-3 : Baghdad bomb kills five US troops
English-4 : Baghdad Bomber Kills Five US Troops
English?? : Bomber suicide slaughters five US soldiers Baghdad

It can be observed the last one is wierd to be called an English sentence, though it has English words and some phrases. Notice that other four sentences are in good English. Now, if we have to analyse the sentence and pin down precisely what went wrong with it we can identify at least two problems: i) the English words are not in the *right* order and ii) is *slaughters* the right word in this context?
Now from the perspective of language, if we leave aside the aspects of style (simply we do not want the machine to be Shakesphere, at least not as yet), exactly these are the two aspects for the machine to focus on. The first issue corresponds to what is called *language modeling*, while the second one can be understood as *content-transfer* (or in slightly subtle way as *adequecy* - am i adequetly capturing the content of the source language). In other way, we can ask ourself given the source sentence am i conveying the same meaning in the target language given constraints in it.

9.2 Translation Process

The process of translation can be seen in a slightly unorthodox way (compared with how humans translate) as having four steps as below (assuming that we are translating from English to French).

Step-1: For each word in English, determine the number of words in the French translation *produced* by this word; this is called the Fertility of the word.

Mary did not slap the green witch

Here the word 'did' has 0 fertility and hence does not contribute to any word in French translation. In contrast to this, the word 'slap' has the fertility of 3 and hence results in 3 words in French sentence. All the other words have fertility 1.
We should note that while this model allows one source word to produce many target words (one-to-many), this doesn't let many-to-one or many-to-many cases. This is an inherent weakness of this model, which other SMT systems overcome using different ideas.

Step-2: There might be some words present in the target sentence which did not have an equivalent in the source. We assume that these words are generated from a 'Null' word in the source string and are inserted in an appropriate position in the target. For example the French translation for the above sentence will have a new word 'a' to produce the quivalent of 'the' as *a la*. This is referred to as the null-insertion probability, i.e. probability of inserting a new word in the target.

Step-3: Translate each English word to French equivalent(s) taking into consideration their fertility values determined in the first step. Thus the French sentence would now look like (the numbers in the square brackets indicate the index of corresponding English word; Null word is assumed to be in index 0).

Maria[1] no[3] daba[4] una[4] bofetada[4] a[0] la[5] verde[6] burja[7]

This means that we'll need a good dictionary that will contain all the English words and their French equivalents along with their probabilities. From a computational viewpoint, we note that this is nothing but a *translation model* between English and French, which will give us the French equivalents given English words along with their probabilities. We need the probabilities because we might have to choose between different translations for a given word depending on some other constraints.

Step-4: Now, we realise that this is a crude translation and to get a *good* translation according to French grammar, we may have to *distort* the positions of certain words. With the distortion step (aka reordering) applied, the translation would be:

Maria[1] no[3] daba[4] una[4] bofetada[4] a[0] la[5] burja[7] verda[6]

This step indicates the need for *language models* and we should note that this improves the readability (fluency) of the target sentence.

9.3 Evolution of Statistical MT

Before proceeding further, we will briefly see the evolution of statistical machine translation (for a short overview on other approaches to machine translation, see the Sidewalk in later section), which can be summarized as below.

Word-based approaches: Started with Brown et al.[1] though the research began in mid 80s. This models the translation process as translation of individual words in source sentence, accounting for one-to-many alignments and different word orders in target language. Brown et al.[1] presents five different models (called IBM models) for word alignment in increasing levels of sophistication.
The following example (slightly modified from Philip koehn's tutorial [3]) illustrates the idea. The numbers in the square brackets in the target sentence correspond to the source word indices that they are aligned to.

Mary did not slap the green witch

Maria[1] no[3] daba[4] una[4] bofetada[4] a[0] la[5] burja[7] verda[6]

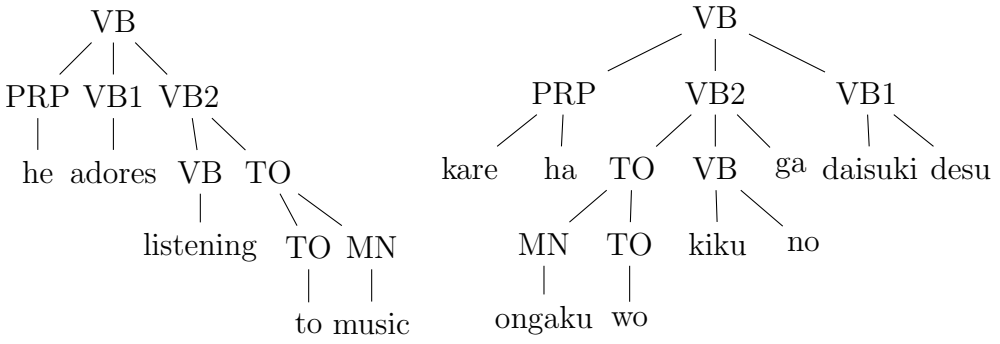
Phrase-based approaches: Beginning in late 90s, these models improved upon the word alignment proposed by IBM models to capture the alignment between phrases and use them as base unit for translation. Here we should note that the phrases do not have any linguistic significance, but are just assumed to be a sequence of words. The example here shows the words grouped together in phrases in both source and target.

(Morgen) (fliege) (ich) (nach Kanada) (zur Konferenz)

(Tomorrow)[1] (I)[3] (will fly)[2] (to the conference)[5] (in Canada)[4]

Syntax-based approaches: Phrase-based models have certain disadvantages like their inability to model discontiguous alignments (*not* in English aligning to *ne . . . pas* in French) and global dependencies in reordering (that appear long distance in either side).

Here, the source or target or both sentences are represented as syntactic trees obtained by parsing and the translation model is based on this. Depending on which side is parsed there are 3 flavours, viz: i) both languages have trees (tree-to-tree model), ii) only target have tree (string-to-tree) and iii) only source have tree (tree-to-string). The following example explains the idea, though it hides the structural transformations during intermediate stages. Refer to the tutorial by [3] for details.
Source: He adores listeening to music.



Target: kare ha ongaku wo kiku no ga daisuki desu

Even though the syntax-based approaches are linguistically appealing they perform poorly when compared with phrase-based systems, which achieve better scores in automatic evaluation in different MT competitions.

9.4 MT Evaluation

Traditional evaluation of MT systems involved humans giving subjective scores for measures such as fluency and adequecy in a pre-determined scale which are then averaged to even out the subjective bias. Lack of consistency in the evaluation process by the same person at different times and between different judges is an obvious issue in this. In addition it is also time consuming and hence is ineffective in keeping track of the performance changes during the system development process.
Automatic techniques of MT evaluation are meant to directly address these issues, i) they will be consistent even when the evaluation scheme is deficient and ii) evaluation can often give some meaningful *score* which can be useful in tracking the system performance. Initial MT evaluation ideas was based

on *word error rate* (WER)- a well-known idea borrowed from the speech processing community.

9.4.1 BLEU Metric

BiLingual Evaluation Understudy- BLEU, proposed by Papineni et al.[4] presented the idea of using the weighted average of n-gram overlaps (for different n) between the translation output and one or more reference translations to assign a score to the translation output. The BLEU score has two components, modified n-gram precision and brevity penalty.
The modified n-gram precision is calculated as the number of n-grams in the candidate that occur in the reference translations (upper bounded by the maximum number of respective n-grams in any of the references) divided by the total n-grams in the candidate. The BLEU is computed for the entire test corpus and not for individual sentences. Mathematically, the modified precision score p_n for the entire test corpus S having different candidates C can be written as:

$$p_n = \frac{\sum_{C \in S} \sum_{n\text{-}gram \in C} C_{max}(n\text{-}gram)}{\sum_{C' \in S} \sum_{n\text{-}gram' \in C'} C(n\text{-}gram')} \tag{9.1}$$

The modified n-gram precision scores are calculated for up to $n = 4$ and are combined by taking their geometric mean with uniform weights.
Candidate translations can now produce shorter translations having some matching n-grams and this will now get higher precision scores. To penalize such shorter translations (as compared with references), a multiplicative *brevity penalty* is introduced as a decaying exponential in r/c and is calculated for entire test corpus with r and c being the lengths of the length of the references and candidate translations and is written as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{9.2}$$

And the BLEU score is written as

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{9.3}$$

BLEU score takes a value between 0 and 1 with higher scores indicating better performance. It has been shown to correlate well with human judgements in several different experiments. Since its publication, BLEU has been widely used for evaluating SMT systems and has given rise to variations such as NIST score (uses arthmetic mean and up to 9-grams).
BLEU has several limitations such as, failure to recognize i) word choice variation and ii) variations in word ordering. Despite these limitations it is being widely used for two important reasons. First it offers a reliable, consistent and simple means of keeping track of the performance variations while requiring just reference translations. Secondly, it serves as an optimization function against which the system parameters can be tuned for better performance (by maximizing BLEU).

9.4.2 Quality of Translation Systems

The tutorial by [3] describe a fun fact about the translation quality between the 11 official languages of EU. 110 systems were trained using the European parliament corpus and BLEU score was computed for each language pair, which resulted in interesting observations. Such as, translating into German and Finnish from any language resulted in lower BLEU and it should be noted that these two languages have complex morphology. Such findings points to potential research directions. Readers are encouraged ro refer to the tutorial for more details.

9.5 Modeling SMT

9.5.1 Noisy-Channel Model

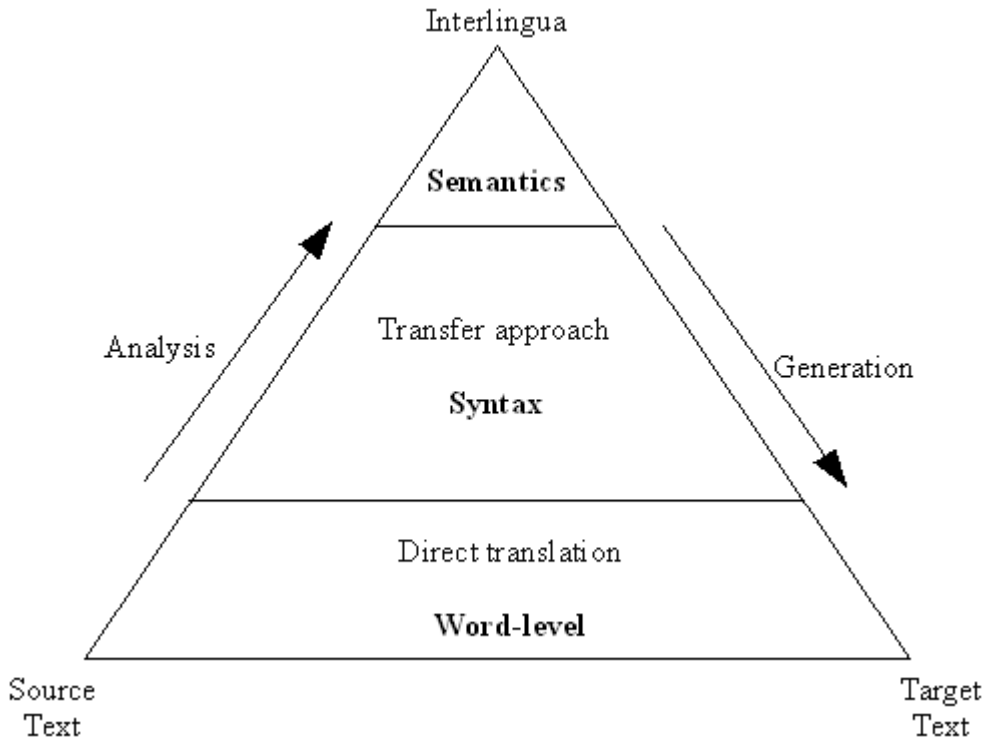
In Statistical MT the translation process is viewed as a noisy-channel model, where the input English sentence is corrupted by some noise in the channel and output as some other foreign language sentence. And the idea is to recover the original English sentence from the foreign language sentence. Probabilistically, this can be expressed using the Bayes rule as:

$$P(e|f) = \arg\max_{e} P(e) * P(f|e) \tag{9.4}$$

The two probability terms on the right side of the equation correspond to whare are called the language model and the translation model respectively.

Sidewalk: MT Hierarchy

Several approaches has been tried for Machine Translation, from the early days of MT research in 1950s. These aproaches differ from each other at various levels of processing the source and/or target languages. They are popularly illustrated using *MT Pyramid* as shown in the figure below.



At the simplest level the source text can be translated at word-level without any syntactic analysis of either the source or target side and is called *direct translation*. It should be noted that this might involve morphological processing on either or both sides. This might result in a ungrammatical translations if the target language is different from the source language in a considerable way.

Sophistication in the translation process is achieved by incorporating more analysis and/ or generation corresponding to source and target texts. An *transfer-based* MT system will normally be based on syntactic tree transfer from source to target text.
Progressively research was carried out to include semantics (meanings in the language) as well in the translation process with the hope of improving the performance. One popular approach in this direction is called *Interlingua*, where the concepts/meanings of different languages are mapped to a common representation (interlingua). Thus in this framework, the translation between two languages A and B would involve two stages, viz. translation from A to interlingua and translation from interlingua to B. However this proved to be considerably difficult and till date no such successful system (beating the syntactic systems) has been developed. For a popular introduction about MT, refer to the Wikipedia entry [5].

Though the initial idea of Statistical MT was known since 1949 when it was proposed by Warren Weaver, it was not used widely as a practical technique until after [1] proposed *IBM Models*- a series of five models based in increasing sophistication of generative process. We will use English and French respectively as source and target languages in explaining the models following their legacy.
IBM Model-1 is a simple model that generates a French sentence from the English setence, by first deciding the length of the French string assuming all reasonable lengths to be equally likely. Then for each position in the French string it finds the how this should be connected to English string (assuming equal probability for all possible connections) and then what French word to place there. The reader is encouraged to refer [1] for details; here we restrict ourselves to IBM Model-3.

9.5.2 IBM Model-3

Earlier in section 9.2 we defined the translation process by a series of steps; which actually is based on Model-3. We now formalize the process in this section and mention the key mathematical equations related to this model. Model 3 starts by choosing for each word in English string the number of the

French words produced. This is called *fertility* ϕ and it depends only on the specific English word. It should be noted that some English word might have zero fertility indicating that it was not translatd into the French string. After choosing the fertility ϕ_i for each English word e_i, the model now generates $\phi - i$ French words depending only on e_i with some translation probability $t(f|e_i)$.

After generating all the French words, they are permuted to their target position j, based on the position i of the corresponding English word and the lengths l and m of the English and French strings. This is called the *distortion probability* $d(j|i, l, m)$.

Now there are certain words in the French string that do not have corresponding words in original English string. Such, words are assumed to generated from an empty cept in the 0^{th} position of the English string having a fertility ϕ_0 and are placed in the vacant positions with a uniform probability $1/\phi_0$.

We will define two terms now before proceeding further. A list of French words connected to an English word is called a *tablet* and a colletion of tablets is a random variable T called *tableau* of e. T_{ik} is a random variable indicating the k^{th} French word in the i^{th} tablet. The permutation of words is a random variable Π; Π_{ik} represents the random variable for the position in f of th k^{th} word in the i^{th} tablet.

Given an English string e, the probability of finding the French string f and alignment a is given by the equation:

$$P(f, a|e) = \sum_{(\tau,\pi)\in<f,a>} P(\tau, \pi|e) \tag{9.5}$$

where, $P(\tau, \pi|e)$ gives the joint probability of a *tableau*, τ and a permutation, π which can be written as below, explaining the complete generative process.

$$P(\tau,\pi|e) = \prod_{i=1}^{l} P(\phi_i|\phi_1^{i-1},e)P(\phi_0|\phi_1^l,e)* \prod_{i=0}^{l}\prod_{k=1}^{\phi_i} P(\tau_{ik}|\tau_{i1}^{k-1},\tau_0^{i-1},\phi_0^l,e)* \prod_{i=1}^{l}\prod_{k=1}^{\phi_i} P(\pi_{ik}|\pi_{i1}^{k-1},\pi_1^{i-1},\tau_0^l,\phi_0^l,e)* \prod_{k=1}^{\phi_0} P(\pi_{0k}|\pi_{01}^{k-1},\pi_1^l,\tau_0^l,\phi_0^l,e) \quad (9.6)$$

Model 3 makes simplifying assumptions about the fertility, translation and distortion probabilities as mentioned above. Considering these assumptions and summing over all possible alignments $P(f|e)$ can be written as:

$$P(f|e) = \sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l}\phi_i n(\phi_i|e_i)* \prod_{j=1}^{m} t(f_j|e_{a_j}) d(j|a_j,m,l) \quad (9.7)$$

The parameter values for the fertility, translation and distortion probabilities are estimated from a large parallel corpus.

Bibliography

[1] Peter E Brown, Vincent J. Della Pietra, Stephen A. Della Pietra and Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. In Computational Linguistics, volume 19,number 2, June 1993.

[2] Kevin Knight. 1999. *A Statistica MT Tutorial Workbook*. Prepared for JHU Summer Workshop 1999.

[3] Philip Koehn. 2006. *Statistical Machine Translation: the basic, the novel and the speculative*. Tutorial at EACL 2006.

[4] Kishore Papineni, Salim Roukos, Todd Ward and Wi-Jing Zhu. 2002. *BLEU: A Method for Automatic Evaluation of Machine Translation*. In the Proceedings of ACL.

[5] Wikipedia MT page: http://en.wikipedia.org/wiki/Machine_translation.