The decoder in statistical machine translation: how does it work?

Alexandre Patry

RALI/DIRO Université de Montréal

June 20, 2006
The goal of machine translation is the creation of a system that will translate a document without human intervention.
The goal of machine translation is the creation of a system that will translate a document without human intervention.

Some paradigms have been proposed to resolve this problem:

**Symbolic translation** Human experts encode their knowledge in the system.
The goal of machine translation is the creation of a system that will translate a document without human intervention.

Some paradigms have been proposed to resolve this problem:

**Symbolic translation** Human experts encode their knowledge in the system.

**Example-based translation** Knowledge is acquired from a bilingual text (bitext) using basic statistics (similar to learning by analogy).
The goal of machine translation is the creation of a system that will translate a document without human intervention.

Some paradigms have been proposed to resolve this problem:

**Symbolic translation**  Human experts encode their knowledge in the system.

**Example-based translation**  Knowledge is acquired from a bilingual text (bitext) using basic statistics (similar to learning by analogy).

**Statistical machine translation**  Knowledge is acquired on a bitext using statistics.
In statistical machine translation, we try to resolve two problems:

**Modeling** Acquisition and type of knowledges.

**Decoding** Usage of the knowledge to translate a new document.
In statistical machine translation, we try to resolve two problems:

**Modeling** Acquisition and type of knowledges.

**Decoding** Usage of the knowledge to translate a new document.

This presentation focuses on the second problem, the one addressed by the decoder.
Overview

1. The traveler’s decoder
2. Conceptual framework
3. MOOD
4. Implementing a phrase-based decoder
5. Experiments
6. Conclusion
Little story

A French speaking traveler equipped with a bilingual dictionary enters in a New-York store. While reviewing the price chart, he encounters a line that he does not understand:

A sheet of paper .. .. .. .. .. 0.25$

We will look at a process this traveler could use to decode this strange sentence.
Sentence to translate  *A sheet of paper*

<table>
<thead>
<tr>
<th>source</th>
<th>target</th>
<th>source</th>
<th>target</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Un</td>
<td>of</td>
<td>de</td>
</tr>
<tr>
<td>A</td>
<td>Une</td>
<td>of</td>
<td>du</td>
</tr>
<tr>
<td>sheet</td>
<td>feuille</td>
<td>paper</td>
<td>papier</td>
</tr>
<tr>
<td>sheet</td>
<td>drap</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Bilingual dictionary**

**The traveler’s common sense**  The traveller can intuitively evaluate the likelihood of a french sentence.
With these resources in hand, the traveler can use the following algorithm to translate the sentence one word at a time:

1. initialize the set of candidate translations $\mathcal{H}$ with an empty sentence.
2. while there are incomplete sentences in $\mathcal{H}$
   1. pick the least completed translation $h$ from $\mathcal{H}$
   2. for each possible translation $\delta$ for the next word to translate in $h$
      1. append $\delta$ at the end of $h$ and store the result in $h_{copy}$
      2. if $h_{copy}$ is likely following the traveler’s intuition, add it to $\mathcal{H}$
3. return the best candidate in $\mathcal{H}$
The traveler concludes that the most likely translation is "Une feuille de papier."
The traveler concludes that the most likely translation is *Une feuille de papier*.
A sentence containing 10 words having each 5 translations can be translated by more than 9 millions target sentences and the corresponding search graph have more than 12 millions vertices.

\[\text{translations} = 5^{10} \quad \text{vertices} = \sum_{i=0}^{10} 5^i\]
A sentence containing \(10\) words having each \(5\) translations can be translated by more than \(9\) millions target sentences and the corresponding search graph have more than \(12\) millions vertices.

\[
\text{translations} = 5^{10} \quad \text{vertices} = \sum_{i=0}^{10} 5^i
\]

If we allow word reordering, the same sentence will have more than \(35,000\) billions translations and its search graph will contain more than \(43,000\) billions vertices.

\[
\text{translations} = 10!5^{10} \quad \text{vertices} = \sum_{i=0}^{10} \binom{10}{i} i!5^i
\]
A decoder searches the target document $t^*$ having the highest probability to translate a given source document $s$:

$$t^* = \arg \max_{t \in \mathcal{T}} \Pr(t|s)$$
A decoder searches the target document $t^*$ having the highest probability to translate a given source document $s$:

$$t^* = \arg \max_{t \in T} \Pr(t|s)$$

This equation is hard to resolve, it requires all possible target documents to be evaluated!
Some Greek

\[ t^* = \arg\max_{t \in T} \Pr(t|s) \]

Can we translate a source document one step at a time?
Can we translate a source document one step at a time?

\[ t^* = \arg \max_{t \in T} \Pr(t|s) \]

where \( \Delta(s,t) \) is a set containing all the sequences of transformations that can be applied to an initial target sentence to translate \( s \) by \( t \).
\[ t^* = \arg\max_{t \in T} \sum_{\delta^n_1 \in \Delta(s,t)} \Pr(t, \delta^n_1 | s) \]

This equation is still hard to resolve, we thus redefine the problem:

\[ \hat{t} = \arg\max_{t \in T} \max_{\delta^n_1 \in \Delta(s,t)} \Pr(t, \delta^n_1 | s) \]
More Greek

\[ t^* = \arg\max_{t \in T} \sum_{\delta^n_1 \in \Delta(s, t)} \Pr(t, \delta^n_1 | s) \]

This equation is still hard to resolve, we thus redefine the problem:

\[ \hat{t} = \arg\max_{t \in T} \max_{\delta^n_1 \in \Delta(s, t)} \Pr(t, \delta^n_1 | s) \]
The more likely translation is *Une feuille de papier* (\(0.1 + 0.35 = 0.45 > 0.4\)).
The more likely translation is *Une feuille de papier* 
\((0.1 + 0.35 = 0.45 > 0.4)\).

The more likely sequence of transformations leads to the target sentence *Un drap de papier* 
\((0.4 > 0.2 \text{ and } 0.4 > 0.35)\).

We can’t win all the time!
Most decoders assume that the sentences of a document are independents one from each others. The decoder can thus translate each sentence individually.

Shortcomings:

- A sentence cannot be omitted, merged with another one, repositioned or sliced by the decoder.
- The context of a sentence is not considered when it is translated.
The decoder’s task

The task of the decoder is to use its knowledge and a density function to find the best sequence of transformations that can be applied to an initial target sentence to translate a given source sentence.
The decoder’s task

The task of the decoder is to use its knowledge and a density function to find the best sequence of transformations that can be applied to an initial target sentence to translate a given source sentence.

This problem can be reformulated as a classic AI problem: searching for the shortest path in an implicit graph.
Challenges

Two independent problems must be resolved in order to build a decoder:

**Model representation**  The model defines what a transformation is and how to evaluate the quality of a translation.

**Search space exploration**  Enumerating all possible sequences of transformations is often impracticable, we must smartly select the ones that will be evaluated.
The **partial translation** is a translation that is being transformed. It is composed of:

- the source sentence
- the target sentence that is being built
- a progress indicator that shows how to continue this translation

The source and target sentences can be sequences of words, trees, non-contiguous sentences, ...
The partial translation is a translation that is being transformed. It is composed of:

- the source sentence
- the target sentence that is being built
- a progress indicator that shows how to continue this translation

The source and target sentences can be sequences of words, trees, non-contiguous sentences, . . .

Example

In the traveler’s decoder, a partial translation could be:

- source: A sheet of paper
- target: Une feuille
- progress indicator: The next word to translate is the third one.
A partial translation evolves when a transformation is applied to it. A transformation can take many forms:

- translation of one word
- translation of many words
- reordering of the children of a node in a tree
A partial translation evolves when a transformation is applied to it. A transformation can take many forms:

- translation of one word
- translation of many words
- reordering of the children of a node in a tree

Example

In the traveler’s decoder, an example transformation could be:

Add the word *feuille* at the end of the target sentence and update the progress indicator.
A cost quantify the quality of a partial translation. Usually, it evaluates at least:

- the likelihood of the transformations
- the fluency of the target sentence generated so far
- the word reordering that occurred

The cost is used to identify the partial translations to dismiss and to select the best complete translation when the search ends.
A cost quantify the quality of a partial translation. Usually, it evaluates at least:

- the likelihood of the transformations
- the fluency of the target sentence generated so far
- the word reordering that occurred

The cost is used to identify the partial translations to dismiss and to select the best complete translation when the search ends.

**Example**

In the traveler’s decoder, the cost was the common sens of the traveler. It allowed him to dismiss unlikely partial translations like *Un feuille* and to prefer *Une feuille de papier* to *Un drap de papier*.
The transformation generator takes as input a partial translation and outputs the set of transformations that can be applied to it.
The transformation generator takes as input a partial translation and outputs the set of transformations that can be applied to it.

Example

In the traveler’s decoder, the transformation generator indicates that the partial translation *Une feuille* can be transformed to *Une feuille du* or *Une feuille de*. 
A hypothesis is made of a partial translation and of a cost.
The task of the search strategy is two-fold:

- Deciding the order in which the hypotheses are explored.
- Identify the hypotheses to dismiss (using the value of the cost).

Example

In the traveler’s decoder, the search strategy was a breadth-first search where the unlikely hypotheses were dismissed.
Putting it all together

- Each vertex is a hypothesis (partial translation and cost)
- Each edge corresponds to a transformation.
- The transformation generator enumerates the out-edges of each vertex.
- The search strategy defines how to explore the graph.
What is MOOD?

- An acronym for *Modular Object-Oriented Decoder*.
- An architecture decomposing a decoder in six reusable modules.
- A C++ object-oriented framework to create decoders.
- A project that is freely available (as in speech and as in beer) under the GPL license.
What is MOOD?

- An acronym for *Modular Object-Oriented Decoder*.
- An architecture decomposing a decoder in six reusable modules.
- A C++ object-oriented framework to create decoders.
- A project that is freely available (as in speech and as in beer) under the GPL license.

Why MOOD?

- To ease the development of new decoders.
- To give us a tool for research in statistical machine translation.
To see if MOOD can be used with success, we used it to create RAMSES, a new implementation of PHARAOH (Koehn, 2004), a popular state of the art decoder.

PHARAOH uses a phrase-based model, one transformation can translate a sequence of contiguous words.
To see if MOOD can be used with success, we used it to create RAMSES, a new implementation of PHARAOH (Koehn, 2004), a popular state of the art decoder.

PHARAOH uses a phrase-based model, one transformation can translate a sequence of contiguous words.

**Example**

A phrase-based model can have rules like:

- *red herring* → *distraction*
- *house of commons* → *chambre des communes*
A partial translation is made of:

- **source** a sequence of words
- **target** a sequence of words
- **progress indicator** a mask indicating the words that have been translated so far and the position of the next word to translate for the translation to be monotone.

Example

<table>
<thead>
<tr>
<th>source</th>
<th>what a wonderful world</th>
</tr>
</thead>
<tbody>
<tr>
<td>progress</td>
<td>1101, next=5</td>
</tr>
<tr>
<td>target</td>
<td>quel monde</td>
</tr>
</tbody>
</table>
A transformation is composed of a rule and of the position at which this rule applies.

A rule translates a sequence of source words by a sequence of target words with a certain probability.

Example

<table>
<thead>
<tr>
<th>rule</th>
<th>what a → quel with probability 0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>position</td>
<td>1</td>
</tr>
</tbody>
</table>
The cost that is used by RAMSES is a weighted sum of:

Sum of the log-probability of the rules applied so far Evaluates the likelihood of the transformation sequence.
The cost that is used by RAMSES is a weighted sum of:

**Sum of the log-probability of the rules applied so far** Evaluates the likelihood of the transformation sequence.

**Language model** Evaluates the fluency of the target sentence.
The cost that is used by RAMSES is a weighted sum of:

**Sum of the log-probability of the rules applied so far** Evaluates the likelihood of the transformation sequence.

**Language model** Evaluates the fluency of the target sentence.

**Distortion** Penalize the word reordering that takes place between the source and the target sentence.
The cost that is used by RAMSES is a weighted sum of:

**Sum of the log-probability of the rules applied so far** Evaluates the likelihood of the transformation sequence.

**Language model** Evaluates the fluency of the target sentence.

**Distortion** Penalize the word reordering that takes place between the source and the target sentence.

**Length penalty** Control the length of the generated target sentences.
The cost that is used by RAMSES is a weighted sum of:

**Sum of the log-probability of the rules applied so far** Evaluates the likelihood of the transformation sequence.

**Language model** Evaluates the fluency of the target sentence.

**Distortion** Penalize the word reordering that takes place between the source and the target sentence.

**Length penalty** Control the length of the generated target sentences.

**Heuristic** Estimates the cost needed to complete the translation.
The transformation generator returns all the transformations that translates a sequence of words that have not been already translated.

We can restrict the search space by limiting the number of source words that can be skipped between two consecutive transformations.
Example

With the following partial translation:

- source: what a wonderful world
- progress: 1100, next=3
- target: quel

The transformation generator could return:

- rule: wonderful → merveilleux with probability 0.3
- position: 3
**Example**

With the following partial translation:

- **source**: *what a wonderful world*
- **progress**: 1100, next=3
- **target**: *quel*

The transformation generator could return:

- **rule**: `wonderful` → `merveilleux` with probability 0.3
  - **position**: 3

- **rule**: `wonderful` → `splendide` with probabilité 0.1
  - **position**: 3
Example

With the following partial translation:

source: *what a wonderful world*

progress: 1100, next=3

target: *quel*

The transformation generator could return:

- rule: *wonderful* → *merveilleux* with probability 0.3
  position: 3

- rule: *wonderful* → *splendide* with probabilité 0.1
  position: 3

- rule: *world* → *monde* with probability 0.7
  position: 4
RAMSES—Search strategy

RAMSES uses a beam search strategy:

- There are $N + 1$ stacks of hypotheses (where $N$ is the number of source words).
- An hypothesis where $x$ words are already translated is stored in the $x$th stack.
- Stacks are visited in order.
- Each stack is pruned independently from the others.

\begin{itemize}
  \item H1
  \item Hj
  \item Hk
  \item Hl
  \item Hz
\end{itemize}

0 word translated 1 word translated 2 words translated N words translated
WMT’06 (Koehn and Monz, 06)

- Europarl corpus (Koehn’05)
- 6 translation directions: (fr, es, de) \(\leftrightarrow\) en
- http://www.statmt.org/wmt06

<table>
<thead>
<tr>
<th>corpus</th>
<th>nb. of sentence pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAIN</td>
<td>(\approx 700 000)</td>
</tr>
<tr>
<td>DEV</td>
<td>500 (of 2000)</td>
</tr>
<tr>
<td>TEST</td>
<td>2 000</td>
</tr>
</tbody>
</table>

✓ an open setting for testing new ideas and fairly compare different translation systems
A pairwise comparison of RAMSES and PHARAOH

- Same language and translation models (obtained using SRILM, GIZA++ and the tools available at http://www.statmt.org)
- Same function to maximize: a weighted sum of 8 features

\[ \lambda_{lp} \text{length penalty} + \lambda_{lm} \text{language model} + \lambda_d \text{distortion} + \sum_{i=1}^{5} \lambda_i \text{ith translation table score} \]

- Separate tuning of the 8 coefficients using (Och, 2003) (minimization of the BLEU score on the DEV corpus using a smart grid search)
- Automatic evaluation using BLEU
The translations produced by RAMSES and PHARAOH were evaluated using the BLEU score:

\[
\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right)
\]

\[
\text{BP} = \begin{cases} 
1 & \text{si } c \leq r \\
\exp(1 - r/c) & \text{si } c > r 
\end{cases}
\]

where

- \( c \) is the number of words in the target document
- \( r \) is the number of words in the reference document
- \( p_n \) is the ratio of target \( n \)-grams that are shared with the reference.

The BLEU score is a value between 0 and 1 and a higher score is better.
## BLEU results

<table>
<thead>
<tr>
<th>Décodeur</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>German → English</strong></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>25.15</td>
</tr>
<tr>
<td>RAMSES</td>
<td>24.49</td>
</tr>
<tr>
<td><strong>Spanish → English</strong></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>30.65</td>
</tr>
<tr>
<td>RAMSES</td>
<td>30.48</td>
</tr>
<tr>
<td><strong>French → English</strong></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>30.42</td>
</tr>
<tr>
<td>RAMSES</td>
<td>30.43</td>
</tr>
<tr>
<td><strong>English → allemand</strong></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>18.03</td>
</tr>
<tr>
<td>RAMSES</td>
<td>18.14</td>
</tr>
<tr>
<td><strong>English → Spanish</strong></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>29.40</td>
</tr>
<tr>
<td>RAMSES</td>
<td>28.75</td>
</tr>
<tr>
<td><strong>English → French</strong></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>30.96</td>
</tr>
<tr>
<td>RAMSES</td>
<td>31.79</td>
</tr>
</tbody>
</table>
## BLEU results

<table>
<thead>
<tr>
<th>Décodeur</th>
<th>BLEU</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
<th>BP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>German → English</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>25.15</td>
<td>61.19</td>
<td>31.32</td>
<td>18.53</td>
<td>11.61</td>
<td>0.99</td>
</tr>
<tr>
<td>RAMSES</td>
<td>24.49</td>
<td>61.06</td>
<td>30.75</td>
<td>17.73</td>
<td>10.81</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Spanish → English</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>30.65</td>
<td>64.10</td>
<td>36.52</td>
<td>23.70</td>
<td>15.91</td>
<td>1.00</td>
</tr>
<tr>
<td>RAMSES</td>
<td>30.48</td>
<td>64.08</td>
<td>36.30</td>
<td>23.52</td>
<td>15.76</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>French → English</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>30.42</td>
<td>64.28</td>
<td>36.45</td>
<td>23.39</td>
<td>15.64</td>
<td>1.00</td>
</tr>
<tr>
<td>RAMSES</td>
<td>30.43</td>
<td>64.58</td>
<td>36.59</td>
<td>23.54</td>
<td>15.73</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>English → allemand</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>18.03</td>
<td>52.77</td>
<td>22.70</td>
<td>12.45</td>
<td>7.25</td>
<td>0.99</td>
</tr>
<tr>
<td>RAMSES</td>
<td>18.14</td>
<td>53.38</td>
<td>23.15</td>
<td>12.75</td>
<td>7.47</td>
<td>0.98</td>
</tr>
<tr>
<td><strong>English → Spanish</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>29.40</td>
<td>61.86</td>
<td>35.32</td>
<td>22.77</td>
<td>15.02</td>
<td>1.00</td>
</tr>
<tr>
<td>RAMSES</td>
<td>28.75</td>
<td>62.23</td>
<td>35.03</td>
<td>22.32</td>
<td>14.58</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>English → French</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHARAOH</td>
<td>30.96</td>
<td>61.10</td>
<td>36.56</td>
<td>24.49</td>
<td>16.80</td>
<td>1.00</td>
</tr>
<tr>
<td>RAMSES</td>
<td>31.79</td>
<td>61.57</td>
<td>37.38</td>
<td>25.30</td>
<td>17.53</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Principal results highlighted in WMT’06:

- The baseline phrase-based system is not that far from the best systems.
- The quality of the translations produced by SMT systems clearly drops when translating out-of-domain corpora.
MOOD can be used to create real life decoders.
- MOOD can be used to create real life decoders.
- If the features of PHARAOH suit your needs, then PHARAOH is preferred to RAMSES.
MOOD can be used to create real life decoders.

If the features of PHARAOH suit your needs, then PHARAOH is preferred to RAMSES.

RAMSES and MOOD are good contenders for research.
In brief:

- A decoder search for the best sequence of transformations that translates a source sentence.
In brief:

- A decoder search for the best sequence of transformations that translates a source sentence.
- A decoder can be divided in two independent parts:
In brief:

- A decoder search for the best sequence of transformations that translates a source sentence.
- A decoder can be divided in two independent parts:
  - A model representation (transformations, partial translations, cost and transformation generator)
In brief:

- A decoder search for the best sequence of transformations that translates a source sentence.
- A decoder can be divided in two independent parts:
  - A model representation (transformations, partial translations, cost and transformation generator)
  - A search strategy that defines the order in which the hypotheses are explored and that defines a pruning strategy.
Conclusion

In brief:

- A decoder search for the best sequence of transformations that translates a source sentence.
- A decoder can be divided in two independent parts:
  - A model representation (transformations, partial translations, cost and transformation generator)
  - A search strategy that defines the order in which the hypotheses are explored and that defines a pruning strategy.
- MOOD is a modular open source framework that can be used to implement new decoders.
In brief:

- A decoder search for the best sequence of transformations that translates a source sentence.
- A decoder can be divided in two independent parts:
  - A model representation (transformations, partial translations, cost and transformation generator)
  - A search strategy that defines the order in which the hypotheses are explored and that defines a pruning strategy.
- MOOD is a modular open source framework that can be used to implement new decoders.
- RAMSES provides as good translations as PHARAOH, but is open source.
Future works

Research:

- We can probably do better than a weighted sum.
- See how the context of a sentence can be used.
- Phrase-based models overfit, see if we can do better.

Future work for MOOD:

- Write a programmer manual.
- Add new decoders to MOOD.
- Speed up RAMSES.
The decoder in statistical machine translation: how does it work?

A statistical machine translation system translates a source document by the target document having the highest probability to translate it. Such a system is made of a model and of a decoder. The model computes the probability that a document translates another one and the decoder uses the model to find the target document having the highest probability to translate a source document.

In this presentation, I will explain how a state-of-the-art decoder for a phrase-based model works and I will present MOOD, a framework to develop such a decoder.

Dans le décodeur statistique machine traduction : comment cela se passe-t-il?


Dans cette présentation, je vais vous expliquer comment une pointe d’un décodeur phrase-based modèle fonctionne et je présenterai humeur, un cadre à développer un tel décodeur.
Thank you!

http://smtmood.sourceforge.net