CA446

Statistical Machine Translation

Week 9: Linguistic Knowledge

Lecturer: Qun Liu

Lab Tutor: Xiaofeng Wu, Iacer Calixto

2nd Semester, 2014-2015 Academic Year

http://computing.dcu.ie/~qliu/CA446

Content

- **Stack Decoding**
- More on Decoding…
- Linguistics Knowledge for MT
- Using Linguistic Knowledge in Pre-processing
- Using Linguistic Knowledge during Translation
- Using Linguistic Knowledge in Post-processing
- Exercises

Stack Decoding Algorithm

- Partial translations or hypotheses are grouped into stacks based on how many words in the input sentence they translate.

Stack Decoding Algorithm

- Partial translations or hypotheses are grouped into stacks based on how many words in the input sentence they translate.
- Working left to right through the stacks, the decoder takes a hypothesis from a stack and a phrase from the input sentence and tries to construct a new hypothesis.

Stack Decoding Algorithm

- Partial translations or hypotheses are grouped into stacks based on how many words in the input sentence they translate.
- Working left to right through the stacks, the decoder takes a hypothesis from a stack and a phrase from the input sentence and tries to construct a new hypothesis.
- The output is the highest scoring hypothesis that translates the entire input sentence.

Stack Decoding Algorithm

- Partial translations or hypotheses are grouped into stacks based on how many words in the input sentence they translate.
- Working left to right through the stacks, the decoder takes a hypothesis from a stack and a phrase from the input sentence and tries to construct a new hypothesis.
- The output is the highest scoring hypothesis that translates the entire input sentence.
- Stacks are kept to a manageable size using pruning.

Stack Decoding Pseudo Code

```
1: place empty hypothesis into stack 0
2: for all stacks 0...n−1 do
    3: for all hypotheses in stack do
        4: for all translation options do
            5: if applicable then
                6: create new hypothesis
                7: place in stack
                8: recombine with existing hypothesis if possible
                9: prune stack if too big
            10: end if
        11: end for
    12: end for
13: end for
```

Stack Decoding Example

Input sentence: *maison bleu*

Goal: Translate it into English using the stack decoding algorithm with histogram pruning (maximum number of hypotheses per stack = 3)

Stack Decoding Example

Phrase pairs in training data:

maison – house

maison – home

bleu – blue

bleu – turquoise

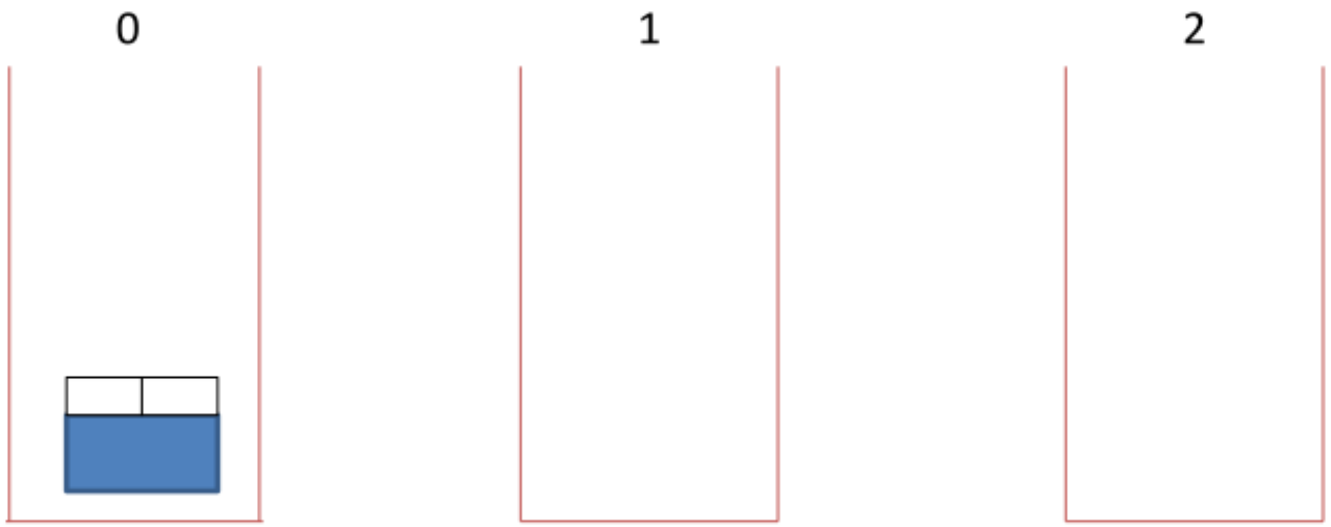
maison bleu – blue house

Stack Decoding Example

All possible translations: *house blue, house turquoise, home blue, home turquoise, blue house, blue home, turquoise home, turquoise house*

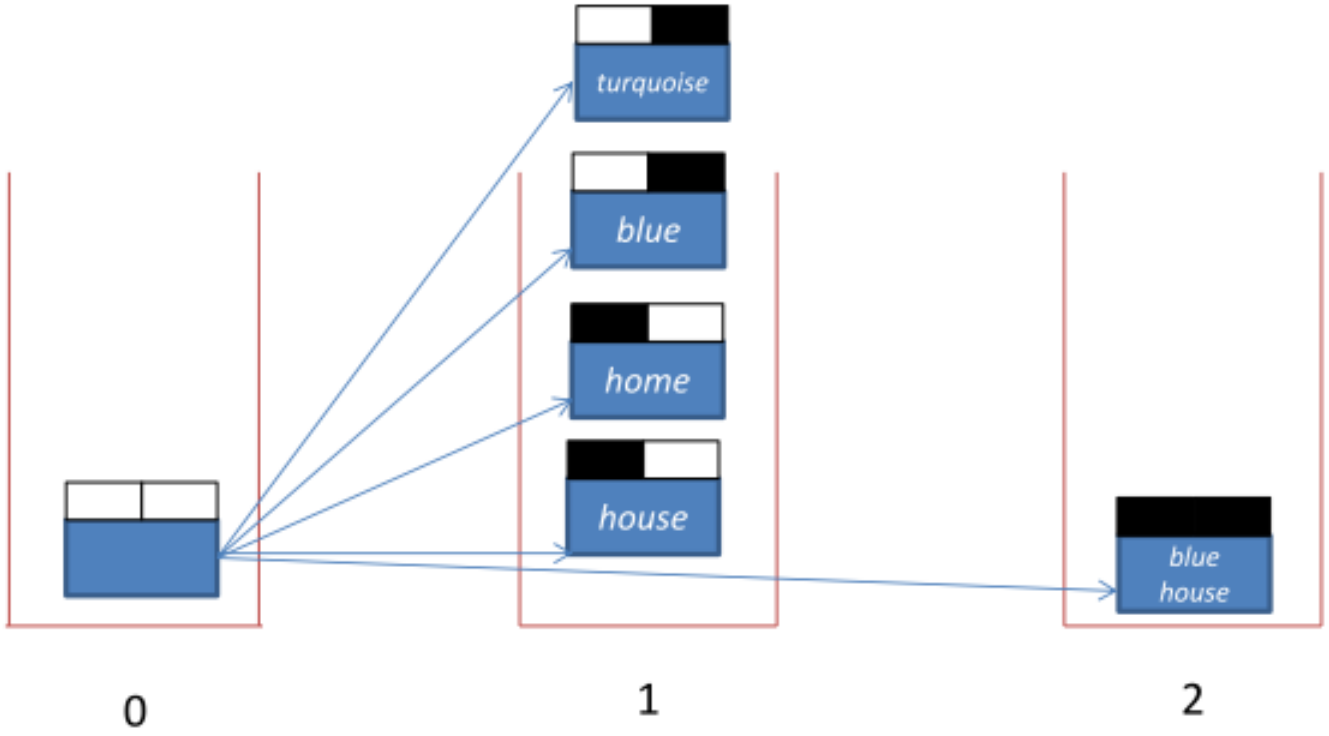
Stack Decoding Example

Insert the empty hypothesis in Stack 0



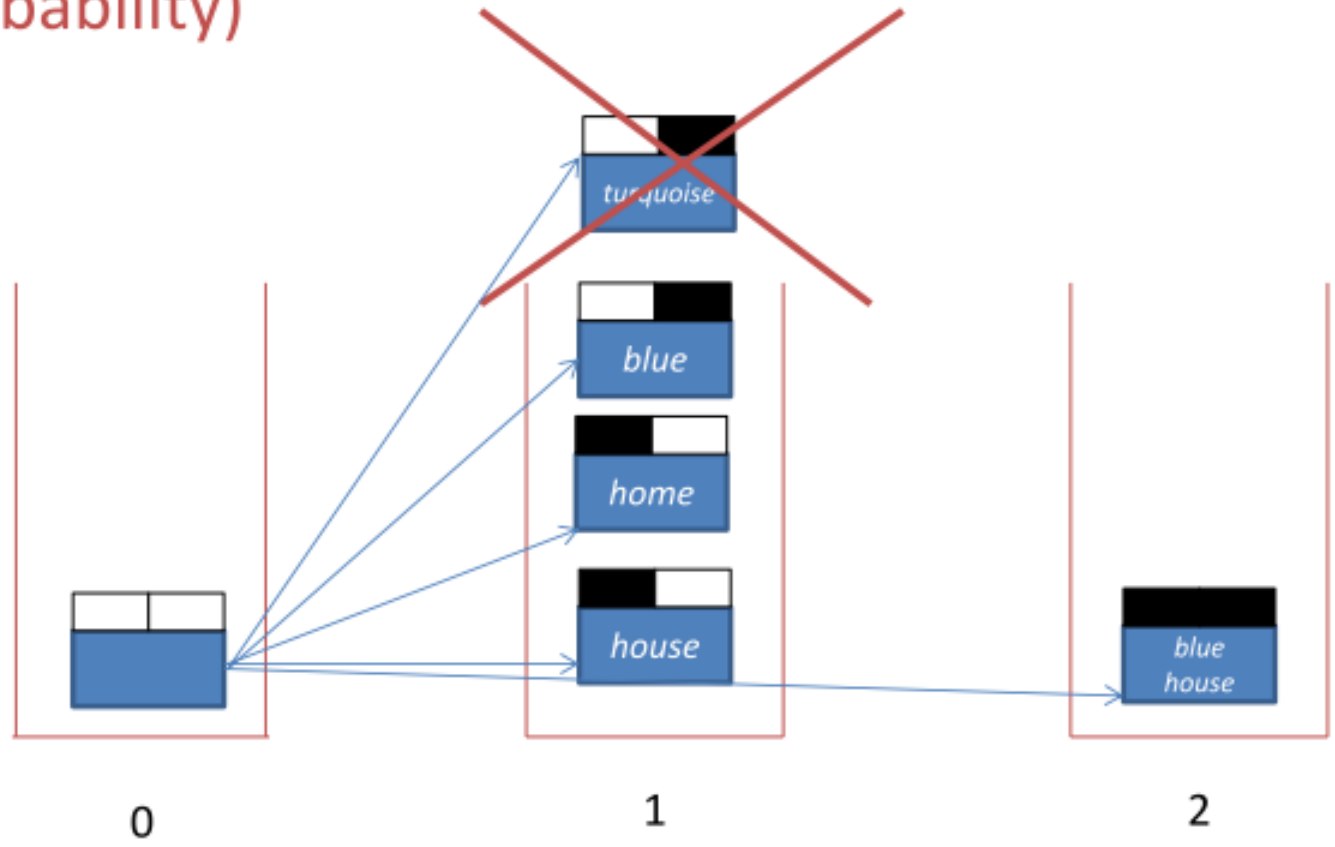
Stack Decoding Example

Construct new hypotheses compatible with empty hypothesis



Stack Decoding Example

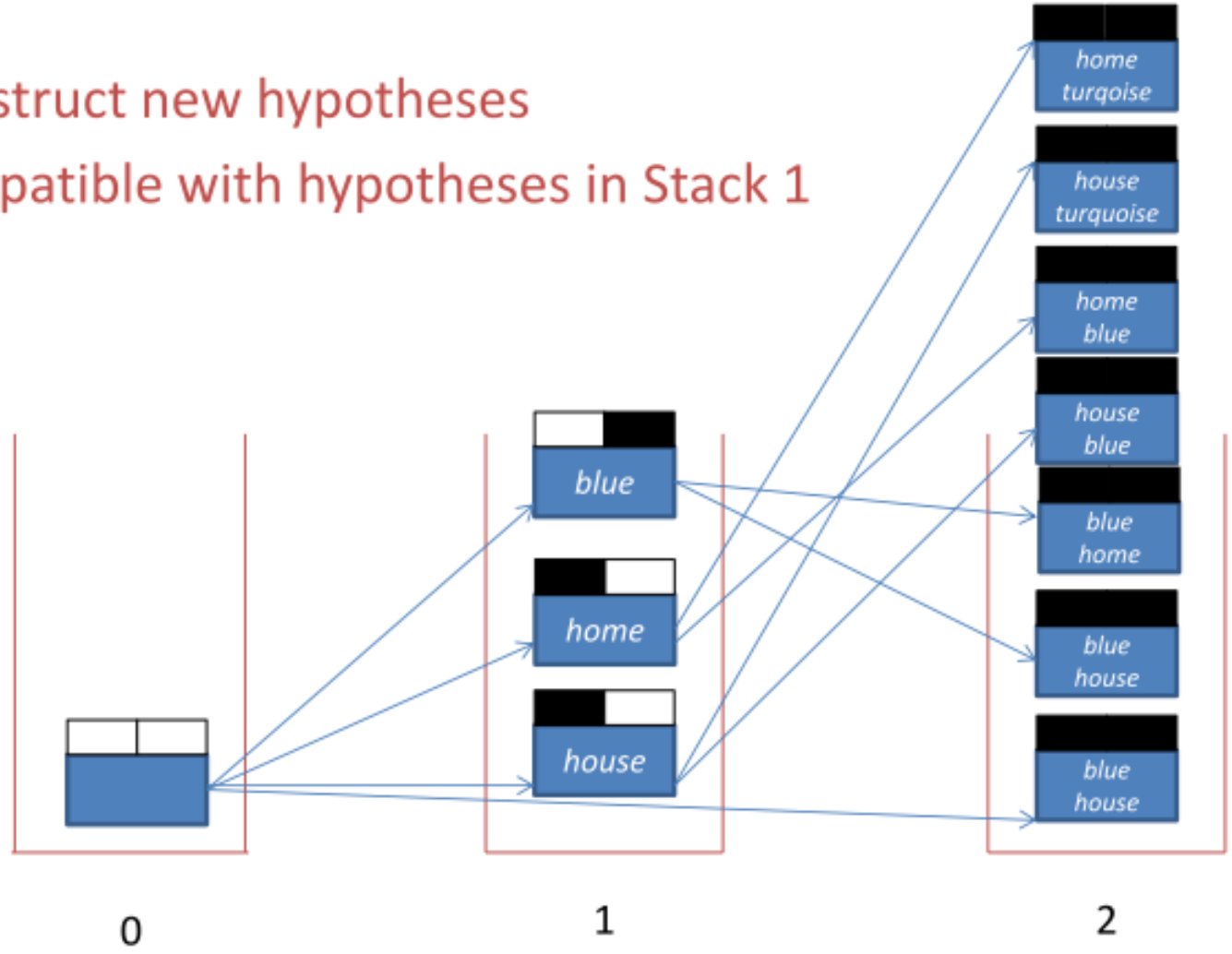
Prune Stack 1 (assume turquoise has lowest probability)



Stack Decoding Example

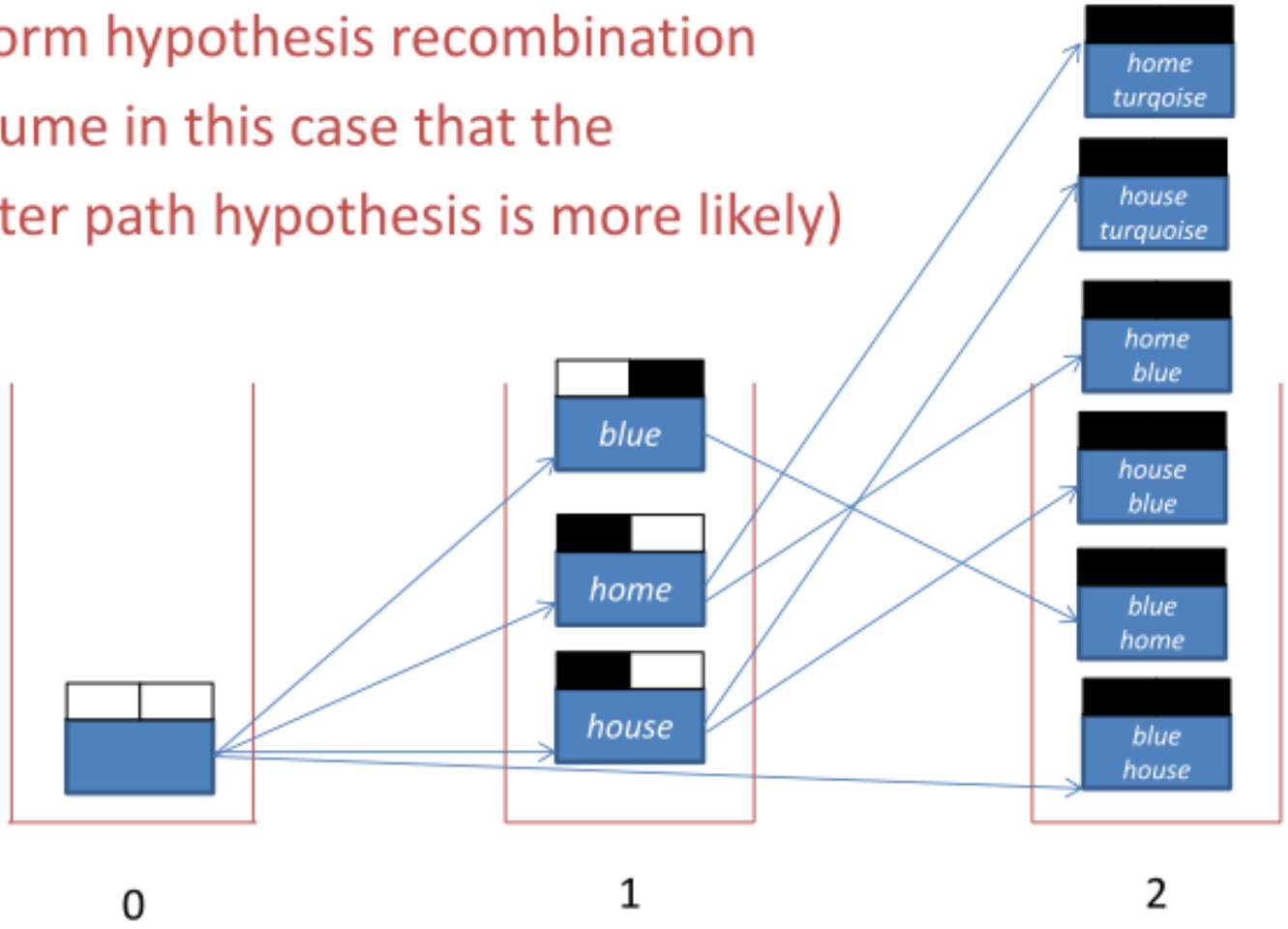
Construct new hypotheses

compatible with hypotheses in Stack 1



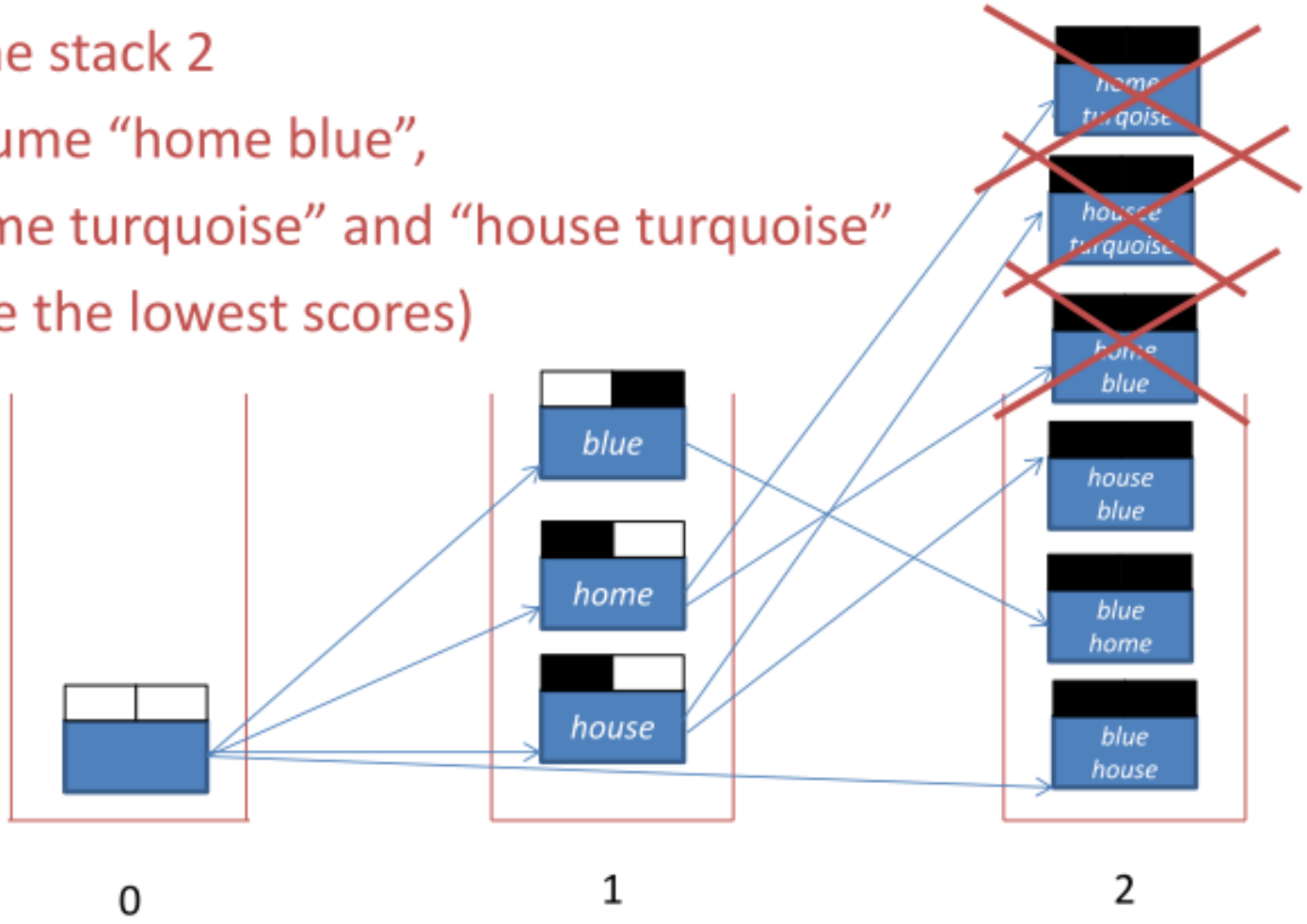
Stack Decoding Example

Perform hypothesis recombination
(assume in this case that the
shorter path hypothesis is more likely)



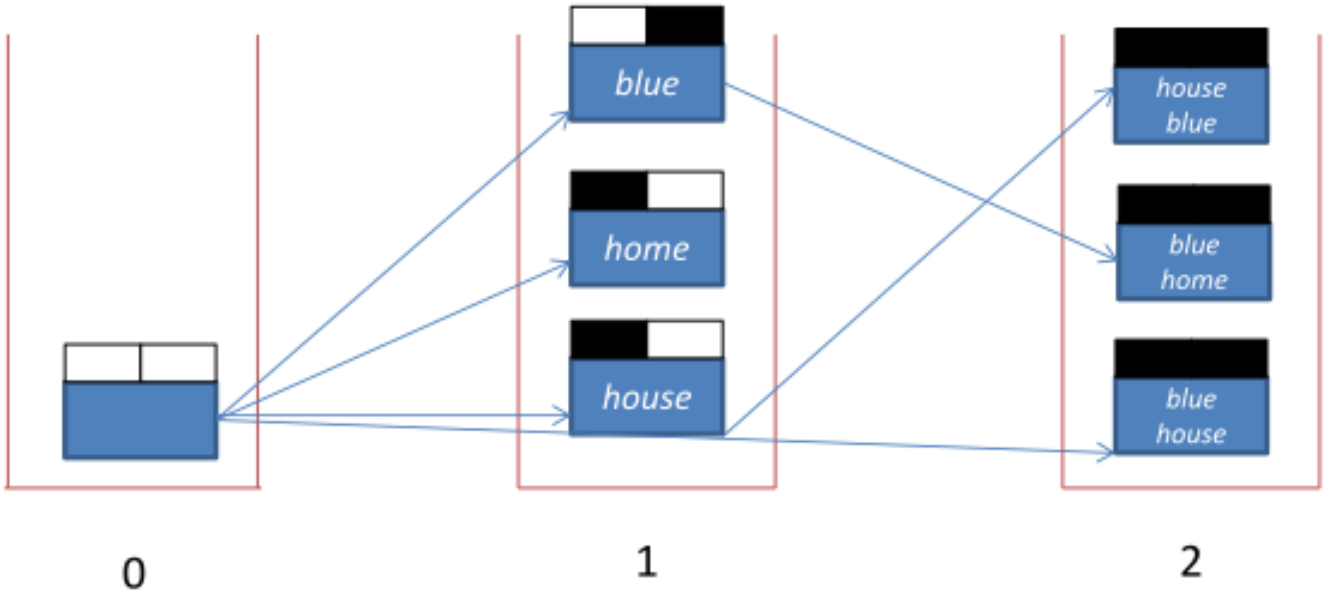
Stack Decoding Example

Prune stack 2
(assume “home blue”,
“home turquoise” and “house turquoise”
have the lowest scores)



Stack Decoding Example

Output most likely translation in Stack 2



Content

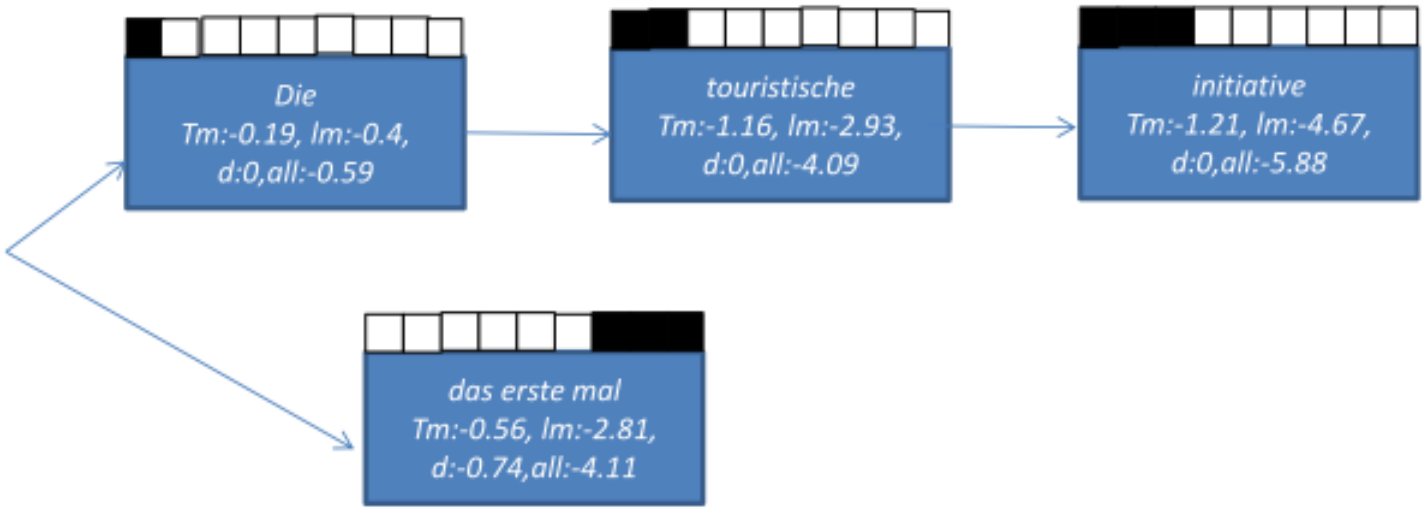
- Stack Decoding
- **More on Decoding…**
  - Linguistics Knowledge for MT
  - Using Linguistic Knowledge in Pre-processing
  - Using Linguistic Knowledge during Translation
  - Using Linguistic Knowledge in Post-processing
- Exercises

Limitation of Stack Decoding Algorithm

Relatively low-scoring hypotheses that are part of a good translation may be pruned….

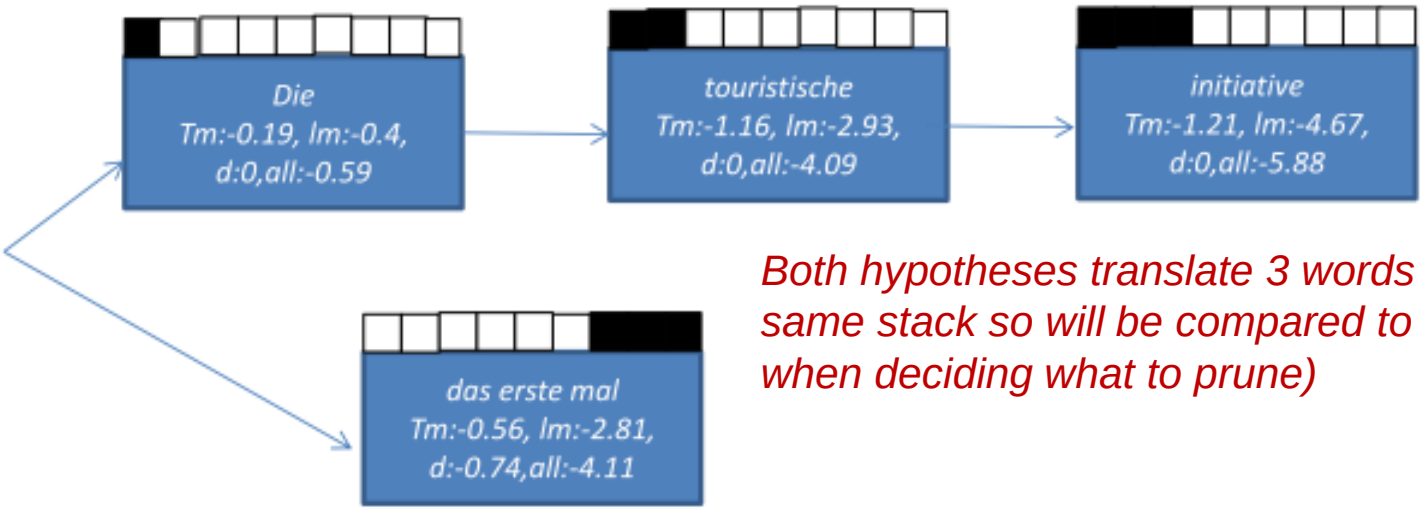
Limitation of Stack Decoding Algorithm

the tourism initiative addresses this for the first time



Limitation of Stack Decoding Algorithm

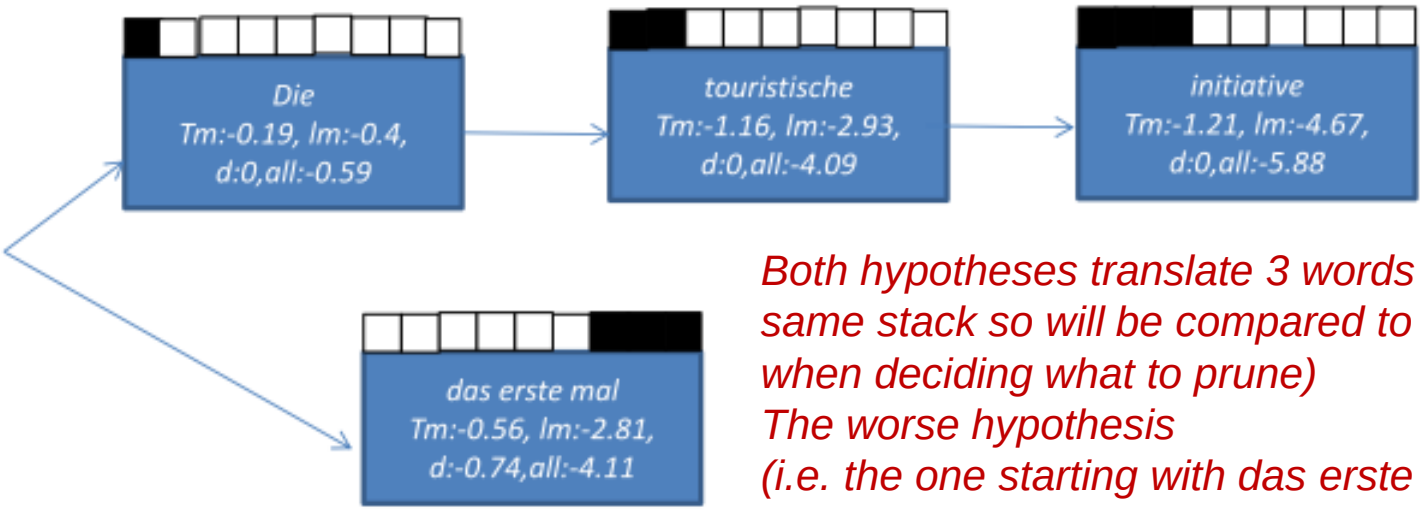
the tourism initiative addresses this for the first time



Both hypotheses translate 3 words (therefore, in the same stack so will be compared to each other when deciding what to prune)

Limitation of Stack Decoding Algorithm

the tourism initiative addresses this for the first time



Both hypotheses translate 3 words (therefore, in the same stack so will be compared to each other when deciding what to prune)
The worse hypothesis (i.e. the one starting with das erste mal) has a better score at this stage because it contains common words that are easier to translate.

Limitation of Stack Decoding Algorithm

How to overcome this limitation?

Limitation of Stack Decoding Algorithm

How to overcome this limitation?

Estimate the future cost associated with a hypothesis by looking at the rest of the input string.

Future Cost

- A hypothesis score is some combination of its probability and its future cost.
- Why is the exact future cost not calculated?

Future Cost

- A hypothesis score is some combination of its probability and its future cost.
- Why is the exact future cost not calculated?

Because calculating the exact future cost of all hypotheses would amount to evaluating the translation cost of all hypotheses – the very thing we use pruning to avoid.

Depth-first versus Breadth-first Search

A distinction is made between decoding algorithms that perform a breadth-first search and those that perform a depth-first search.

Depth-first versus Breadth-first Search

A distinction is made between decoding algorithms that perform a breadth-first search and those that perform a depth-first search.

- A breadth-first search explores many hypotheses in parallel, e.g. stack decoding.
- A depth-first search fully explores a hypothesis before moving on to the next one, e.g. A* decoding.

Search and Model Errors

1. A search error occurs when the decoder misses a translation with a higher probability than the translation it returns.

Search and Model Errors

1. A search error occurs when the decoder misses a translation with a higher probability than the translation it returns.
2. A model error occurs when the model assigns a higher probability to an incorrect translation than to a correct one.

Content

- Stack Decoding
- More on Decoding…
- **Linguistics Knowledge for MT**
- Using Linguistic Knowledge in Pre-processing
- Using Linguistic Knowledge during Translation
- Using Linguistic Knowledge in Post-processing
- Exercises

The story so far…

The SMT approaches we have studied so far are language-independent.

The story so far…

The SMT approaches we have studied so far are language-independent.

Input:

1. A monolingual corpus for training a language model
2. A sentence-aligned parallel corpus for training a translation model

The story so far…

The SMT approaches we have studied so far are language-independent.

Input:

1. A monolingual corpus for training a language model
2. A sentence-aligned parallel corpus for training a translation model

Output:

A translation (of variable quality)

The old way of doing things

Traditional rule-based MT systems tried to translate using hand-coded rules which encode linguistic knowledge or generalisations.

Examples of linguistic knowledge

1. In English, the noun comes after the adjective.
   In French, the noun usually precedes the adjective.

Examples of linguistic knowledge

1. In English, the noun comes after the adjective. In French, the noun usually precedes the adjective.
2. In German, the main verb often occurs at the end of a clause.

Examples of linguistic knowledge

1. In English, the noun comes after the adjective. In French, the noun usually precedes the adjective.
2. In German, the main verb often occurs at the end of a clause.
3. In Irish, prepositions are inflected for number and person.

Examples of linguistic knowledge

1. In English, the noun comes after the adjective. In French, the noun usually precedes the adjective.
2. In German, the main verb often occurs at the end of a clause.
3. In Irish, prepositions are inflected for number and person.
4. In many languages, the subject and verb agree in person and number.

Why bother with linguistic knowledge?

Why would we want to encode linguistic knowledge in an SMT system?

Why bother with linguistic knowledge?

Why would we want to encode linguistic knowledge in an SMT system?

1. Data-driven SMT works well but there is room for improvement (especially for some language pairs).

Why bother with linguistic knowledge?

Why would we want to encode linguistic knowledge in an SMT system?

1. Data-driven SMT works well but there is room for improvement (especially for some language pairs).
2. Data-driven SMT relies on the availability of large corpora. For minority languages, this can be problematic.

Using linguistic knowledge

Where linguistic knowledge might be useful:

Using linguistic knowledge

Where linguistic knowledge might be useful:

1. More global view of fluency than an n-gram language model: *the house is the man is small*

Using linguistic knowledge

Where linguistic knowledge might be useful:

1. More global view of fluency than an n-gram language model: *the house is the man is small*
2. Long-distance dependencies: ***make*** *a tremendous amount of* ***sense***

Using linguistic knowledge

Where linguistic knowledge might be useful:

1. More global view of fluency than an n-gram language model: *the house is the man is small*
2. Long-distance dependencies: ***make*** *a tremendous amount of* ***sense***
3. Word compounding: Knowing that *Aktionsplan* is made up of the words *Aktion* and *plan*

Using linguistic knowledge

Where linguistic knowledge might be useful:

1. More global view of fluency than an n-gram language model: *the house is the man is small*
2. Long-distance dependencies: ***make*** *a tremendous amount of* ***sense***
3. Word compounding: Knowing that *Aktionsplan* is made up of the words *Aktion* and *plan*
4. Better handling of function words: role of *de* in *entreprises de transports* (*haulage companies*)

How to encode linguistic knowledge

How can linguistic knowledge be encoded in an SMT system?

How to encode linguistic knowledge

How can linguistic knowledge be encoded in an SMT system?

1. Before translation (pre-processing)

How to encode linguistic knowledge

How can linguistic knowledge be encoded in an SMT system?

1. Before translation (pre-processing)
2. During translation

How to encode linguistic knowledge

How can linguistic knowledge be encoded in an SMT system?

1. Before translation (pre-processing)
2. During translation
3. After translation (post-processing)

Where do we get this linguistic knowledge?

How do we obtain linguistic knowledge automatically?

Where do we get this linguistic knowledge?

How do we obtain linguistic knowledge automatically?

1. Automatic morphological analysis
2. Part-of-speech tagging
3. Syntactic parsing

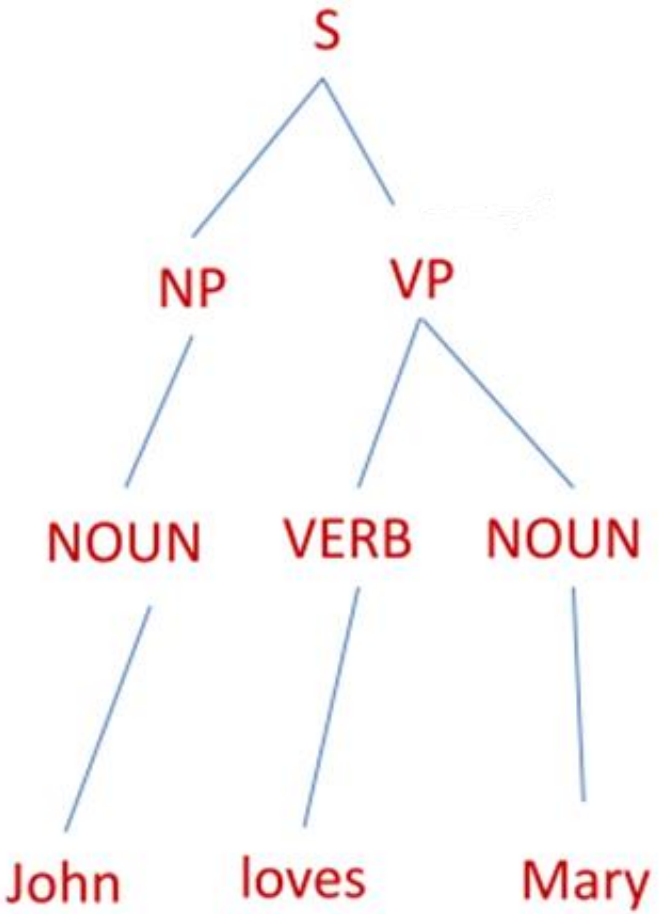
Where do we get this linguistic knowledge?

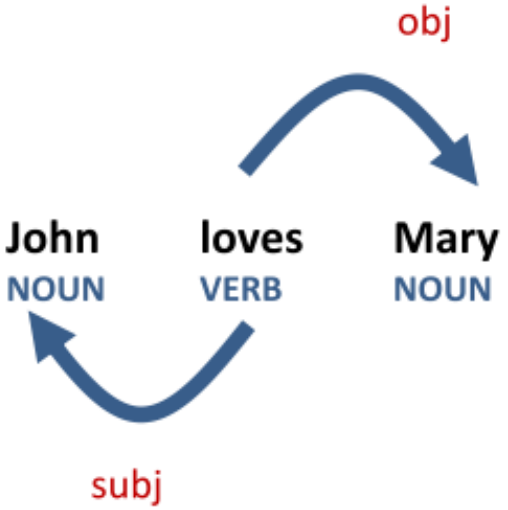
How do we obtain linguistic knowledge automatically?

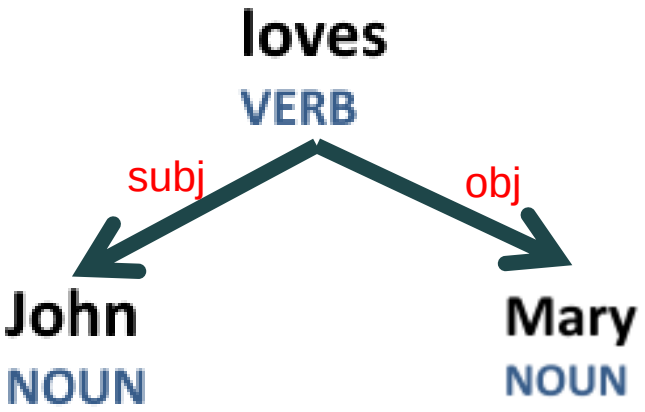
1. Automatic morphological analysis
2. Part-of-speech tagging
3. Syntactic parsing

Can you spot a potential problem here (hint: automatic)?

Syntactic Parsing







Content

- Stack Decoding
- More on Decoding…
- Linguistics Knowledge for MT
- **Using Linguistic Knowledge in Pre-processing**
- Using Linguistic Knowledge during Translation
- Using Linguistic Knowledge in Post-processing
- Exercises

Pre-processing

Encoding linguistic knowledge before translation

1. Word segmentation
2. Word lemmatisation
3. Re-ordering

Word Segmentation

- The notion of what a word is varies from language to language.

Word Segmentation

- The notion of what a word is varies from language to language.
- In some languages, e.g. Chinese, words are not separated by spaces.

Word Segmentation

- The notion of what a word is varies from language to language.
- In some languages, e.g. Chinese, words are not separated by spaces.
- Why is this a problem for machine translation?

Word Segmentation

- The notion of what a word is varies from language to language.
- In some languages, e.g. Chinese, words are not separated by spaces.
- Why is this a problem for machine translation?
- Potential solution: try to perform automatic word segmentation before translation.

Word Segmentation

- The notion of what a word is varies from language to language.
- In some languages, e.g. Chinese, words are not separated by spaces.
- Why is this a problem for machine translation?
- Potential solution: try to perform automatic word segmentation before translation.
- A related challenge is word compounding in German.

Word Lemmatisation

- Languages with a rich inflectional system can have several different variants of the same root form or lemma.

Word Lemmatisation

- Languages with a rich inflectional system can have several different variants of the same root form or lemma.
- Example:

  Latin: *bellum, belli, bello, bella, bellis, bellorum*

  English: war, wars

Word Lemmatisation

- Languages with a rich inflectional system can have several different variants of the same root form or lemma.
- Example:

  Latin: *bellum, belli, bello, bella, bellis, bellorum*

  English: war, wars
- Why is this a problem for translation from an MRL (morphologically rich language) into English?

Word Lemmatisation

- Languages with a rich inflectional system can have several different variants of the same root form or lemma.
- Example:

  Latin: *bellum, belli, bello, bella, bellis, bellorum*

  English: war, wars
- Why is this a problem for translation from an MRL (morphologically rich language) into English?
- Potential solution: before translation, replace full form with lemma

Syntactic Re-ordering

- It is harder to translate language pairs with very different word order.

Syntactic Re-ordering

- It is harder to translate language pairs with very different word order.
- Language model and phrase-based translation model can only do so much.

Syntactic Re-ordering

- It is harder to translate language pairs with very different word order.
- Language model and phrase-based translation model can only do so much.
- Potential solution: transform the source sentence so that its word order more closely resembles the word order of the target language.

Syntactic Re-ordering

- It is harder to translate language pairs with very different word order.
- Language model and phrase-based translation model can only do so much.
- Potential solution: transform the source sentence so that its word order more closely resembles the word order of the target language.
- Consider translating from English to German:

  I walked to the shop yesterday.

  I yesterday to the shop walked.

Content

- Stack Decoding
- More on Decoding…
- Linguistics Knowledge for MT
- Using Linguistic Knowledge in Pre-processing
- **Using Linguistic Knowledge during Translation**
- Using Linguistic Knowledge in Post-processing
- Exercises

During Translation

Encoding linguistic knowledge during translation

During Translation

Encoding linguistic knowledge during translation

1. Log-linear models

During Translation

Encoding linguistic knowledge during translation

1. Log-linear models
2. Tree-based models

During Translation

Encoding linguistic knowledge during translation

1. Log-linear models
2. Tree-based models
3. Syntactic language models

Log-linear models

The log-linear model of translation allows several sources of information to be employed simultaneously during the translation process:

$$p(E|F) = \exp(\sum_{i=1}^{n} \lambda_i h_i(E,F))$$

Log-linear models

The log-linear model of translation allows several sources of information to be employed simultaneously during the translation process:

$$p(E|F) = \exp(\sum_{i=1}^{n} \lambda_i h_i(E,F))$$

Some of this information can be linguistic knowledge, e.g. instead of just calculating $p(ocras|hungry)$ we can also calculate $p(noun|adj)$.

Tree-based models

- A more radical approach is to replace the phrase-based translation model with a different type of translation model entirely.

Tree-based models

- A more radical approach is to replace the phrase-based translation model with a different type of translation model entirely.
- One such model is a tree-based model which is based on the notion of a syntax tree and which allows one, in theory, to better capture structural similarities and divergences between languages.

Tree-based models

- A more radical approach is to replace the phrase-based translation model with a different type of translation model entirely.
- One such model is a tree-based model which is based on the notion of a syntax tree and which allows one, in theory, to better capture structural similarities and divergences between languages.
- During decoding, syntax trees for the source and language pairs are built up simultaneously (or one side only)

Syntactic language models

- N-gram language models are not the only language models.

Syntactic language models

- N-gram language models are not the only language models.
- People have experimented with other kinds of language models, including those based on the notion of a syntax tree.

Syntactic language models

- N-gram language models are not the only language models.
- People have experimented with other kinds of language models, including those based on the notion of a syntax tree.
- Instead of computing the probability of a sentence by multiplying n-gram probabilities, the probability is computed by multiplying the probabilities of the rules that are used to build the syntax tree(s).

Content

- Stack Decoding
- More on Decoding…
- Linguistics Knowledge for MT
- Using Linguistic Knowledge in Pre-processing
- Using Linguistic Knowledge during Translation
- **Using Linguistic Knowledge in Post-processing**
- Exercises

Post-processing

SMT systems return a ranked list of candidate translations.

Linguistic information (of all kinds) can be employed in re-ranking this list using machine learning techniques.

Content

- Stack Decoding
- More on Decoding…
- Linguistics Knowledge for MT
- Using Linguistic Knowledge in Pre-processing
- Using Linguistic Knowledge during Translation
- Using Linguistic Knowledge in Post-processing
- **Exercises**

Discussion

Acknowledgement

Parts of the content of this lecture are taken from previous lectures and presentations given by Jennifer Foster, Declan Groves, Yvette Graham, Kevin Knight, Josef van Genabith, Andy Way.