Elements of DSAI:
Machine Translation

WS 2019/2020

Vera Demberg

Contents for today

- Why is machine translation (MT) difficult?
- Approaches to MT
  - Knowledge-based (rule-based) MT
  - Statistical MT
  - Neural MT
- Evaluation

Contents for today

- **Why is machine translation (MT) difficult?**
- Approaches to MT
  - Knowledge-based (rule-based) MT
  - Statistical MT
  - Neural MT
- Evaluation

Babel Fish, ca. 2007

- Über allen Gipfeln ist Ruh. In allen Wipfeln spürest du kaum einen Hauch
- Over all summits is rest. In all treetops you do not feel breath.
- Über allen Gipfeln ist Rest. In allen Treetops glauben Sie nicht Atem.

Babel Fish, ca. 2007

- Über allen Gipfeln ist Ruh. In allen Wipfeln spürest du kaum einen Hauch
- Over all summits is rest. In all treetops you do not feel breath.
- Über allen Gipfeln ist Rest. In allen Treetops glauben Sie nicht Atem.
- (Google Translate 2020: „Above all summits there is peace. You hardly feel a breath in all the tops.“ „Über allen Gipfeln herrscht Frieden. Sie spüren kaum einen Atemzug in allen Höhen.“

Babel Fish, ca. 2007

- Über allen Gipfeln ist Ruh. In allen Wipfeln spürest du kaum einen Hauch

Offizielle literarische Übersetzung:

- O‘er all the hilltops is quiet now,
  In all the treetops hearest though
  hardly a breath.

Lexical ambiguity

- Homonymy:
  - engl. *rest* → *Rest/Ruhe*

- Polysemy:
  - *breath* → *Atem/Hauch*
  - *Termin* → *appointment / time slot*

- *"gehen"* in Verbmobil (6 of 15 Variants)
  - *Gehen wir ins Theater?* – gehen_move
  - *Gehen wir essen?* – gehen_act
  - *Mir geht es gut*. – gehen_feel
  - *Es geht um einen Vertrag*. – gehen_theme
  - *Das Treffen geht von 3 bis 5*. – gehen_last
  - *Geht es bei Ihnen am Montag?* – gehen_passen

Ambiguity resolution

... through sentence-internal context

- *Wir treffen uns vor dem Frühstück*

  → *before*
- *Wir treffen uns vor dem Hotel*

  → *in front of*

But:

- *Wir treffen uns nach Hamburg*

  → *?*

Ambiguity resolution

… through discourse context

- *Geht es bei Ihnen?*
- *Wo sollen wir uns treffen? Geht das bei Ihnen?*
  → at your place
- *Sollen wir uns am Fünften treffen? Geht das bei Ihnen?*
  → for you

Idioms and collocations

- *Spielkarten geben*

  → *to deal cards*

- *eine Prüfung ablegen*

  → *to take an exam*

- *eine Prüfung abnehmen*

  → *to give an exam*

- *den Fahrschein entwerten*

  → *to validate the ticket*

Language-specific conventional multi-word expressions need to be stored in lexicon as they cannot be derived from rules easily.

Differences in granularity in lexicon

- *I will go to Hamburg tomorrow.*

  → *fahren/fliegen*

- *Ich fahre mit der Bahn nach Hamburg. In Frankfurt muss ich umsteigen.*

  → *change trains*

- *Ich fliege nach Hamburg. In Frankfurt muss ich umsteigen.*

  → *change planes*

Systematic differences in Granularity

- Gender-specific expressions
  - *doctor → Arzt / Ärztin*
  - *teacher → Lehrer / Lehrerin*

- Differences in how tense is used
  - *Ich fahre nach Hamburg → I am going / I will go to Hamburg*

- Differences in verb aspect
  - *Simple Present/ Progressive in English*
  - *Verb aspect in Russian*

Example for granularity differences in Japanese

Deutsch/Englisch → Japanisch

- J: politeness forms
- J: topic marking (new / given)

Japanisch → Deutsch/Englisch

- D: definite / indefinite articles (Japanese doesn‘t have articles)
- J: „Null-Anapher“: Pronouns can be omitted if they can be inferred.

Example

"Termin ausgemacht?"

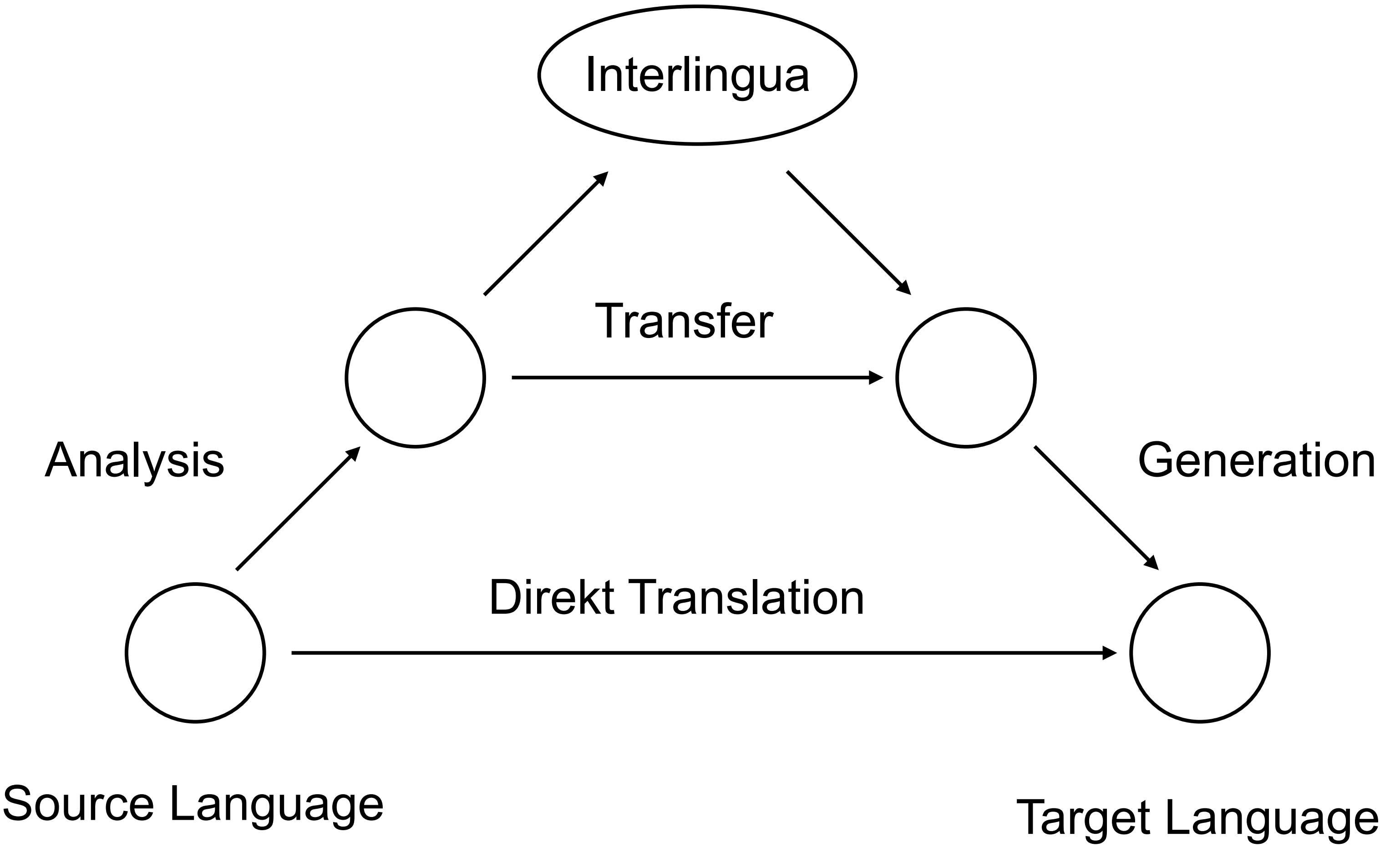
Yotei-wa kimemashita ka.
→ *Hat er (mit Ihnen) einen Termin ausgemacht?*

Go-yotei wa okimeni narimashita ka.
→ *Haben Sie (mit ihm) einen Termin ausgemacht?*

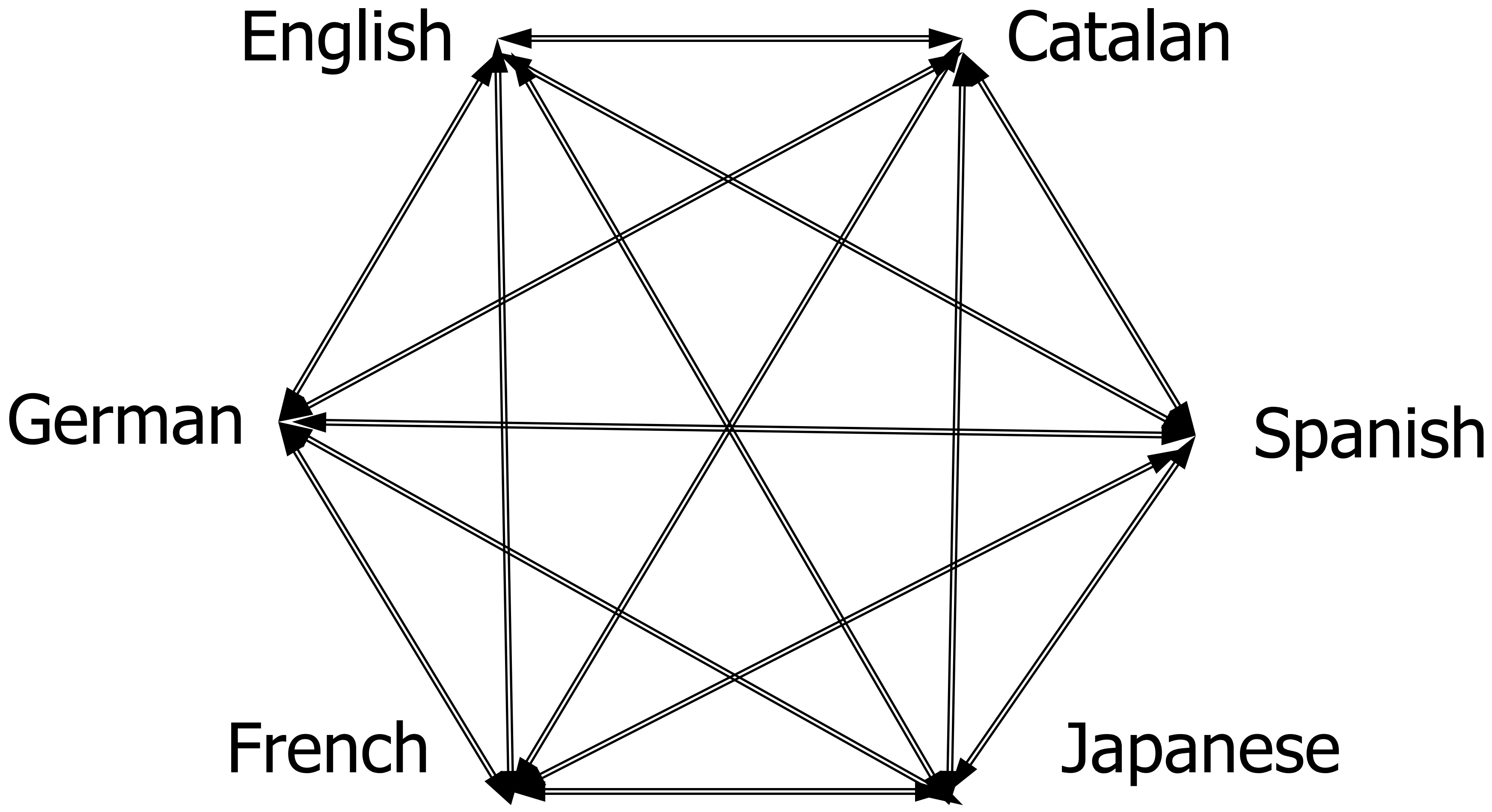
Contents for today

- Why is machine translation (MT) difficult?
- **Approaches to MT**
  - Knowledge-based (rule-based) MT
  - Statistical MT
  - Neural MT
- Evaluation

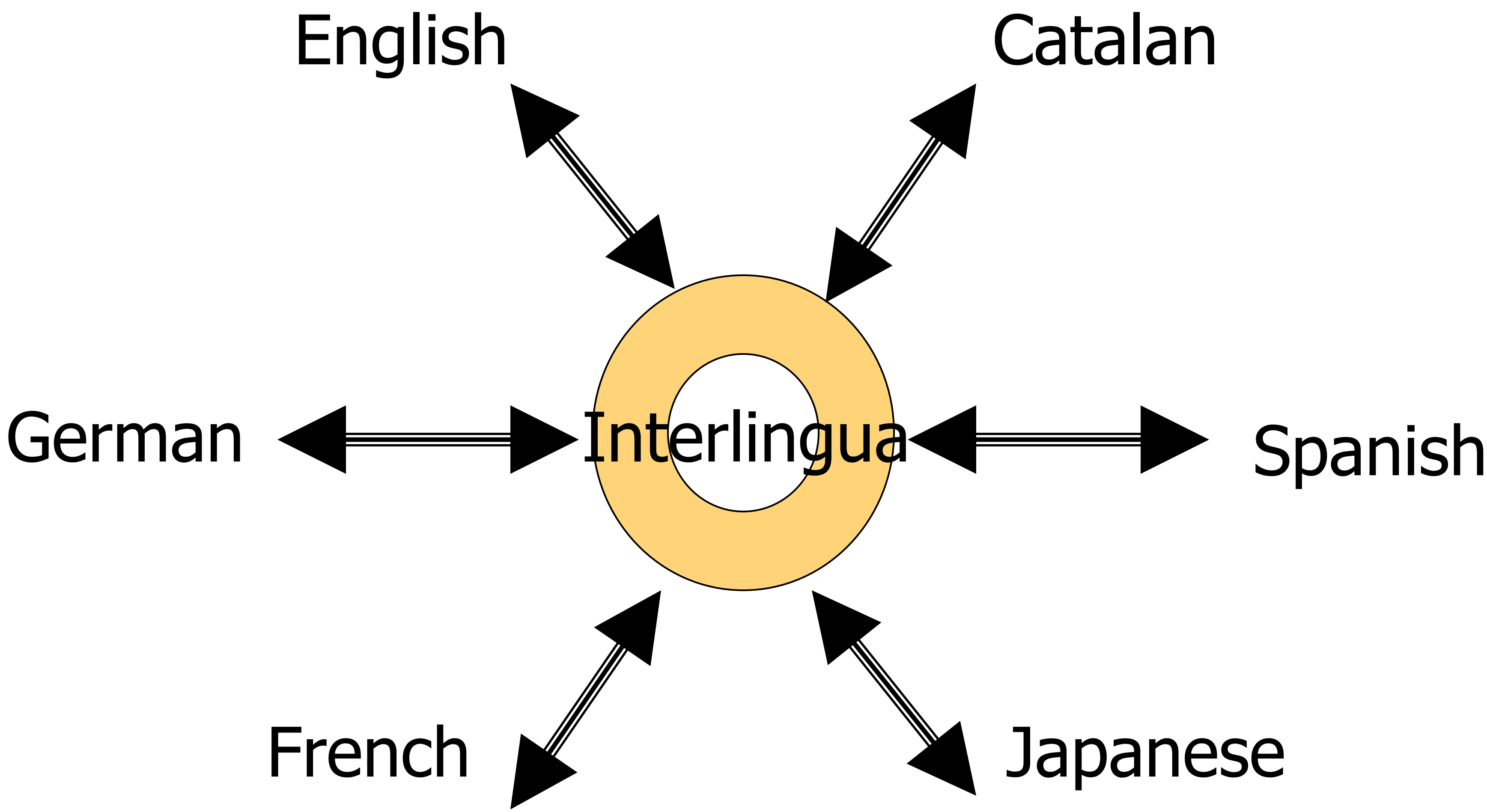
The "Vauquois-Triangle"



Transfer Model



Interlingua Model



Interlingua and Transfer

- For each language, we only need one translation from and to the interlingua.

- In the transfer model, if we add a language, we need to add two translation directions for this language with all other languages in the system.
  - Example: Durch die EU-Erweiterungen 2004 und 2007 wuchsen die offiziellen EU-Sprachen von 11 auf 24 an.
  - Statt 110 Übersetzungspaaren benötigt man 552.

Disadvantage of an Interlingua

- An interlingua must have very fine granularity, i.e. it must be able to represent all distinctions that are linguistically relevant in any of the languages. This means that unnecessary analyses have to be done for related languages.
  - Example: Übersetzung SE-EN benötigt keine detaillierte Bestimmung von Höflichkeitsinformation

Contents for today

- Why is machine translation (MT) difficult?
- Approaches to MT
  - **Knowledge-based (rule-based) MT**
  - Statistical MT
  - Neural MT
- Evaluation

Knowledge-based MT

- Tools: Stemmer/Morphology, Grammar, Lexika for source and target language, transfer rules, language-independet ontologies, world knowledge, inference rules
- Problems
  - coverage: there is a high variety of different syntactic and semantic phenomena and specific translations
  - precision: ambiguity and differences in granularity
- Classical example:
  - SYSTRAN (Babel Fish)
    - E.g. Barbarei -> night club night club egg

Contents for today

- Why is machine translation (MT) difficult?
- Approaches to MT
  - Knowledge-based (rule-based) MT
  - **Statistical MT**
  - Neural MT
- Evaluation

Statistical MT

- We are looking for the most probable sentence in the target language (e.g., in German „D“), given a sentence in the source language (e.g., English „E“).

$$\max_D P(D \mid E)$$

- Hopefully, this reminds you of what we did for speech recognition: instead of a source language, we had a sequence of features extracted from speech, and we were looking for the most likely word sequence.

$$\max_W P(W \mid O)$$

How can we calculate *P(D|E)*?

- Bayes-Rule:

speech recognition:

$$P(W \mid O) = \frac{P(O \mid W) \cdot P(W)}{P(O)}$$

$$\max_W P(W \mid O) = \max_W \frac{P(O \mid W) \cdot P(W)}{P(O)}$$

$$= \max_W P(O \mid W) \cdot P(W)$$

machine translation:

$$P(D \mid E) = \frac{P(E \mid D) \cdot P(D)}{P(E)}$$

$$\max_D P(D \mid E) = \max_D \frac{P(E \mid D) \cdot P(D)}{P(E)}$$

$$= \max_D P(E \mid D) \cdot P(D)$$

Translation model and language model

$$\max_D P(D \mid E) = \max_D P(E \mid D) \cdot P(D)$$

We can determine how „good“ a translation is by observing:

- How well it reflects the input.
  This is approximated by the translation model *P(E|D).*
- The fluency and correctness of the sentence in the target language.
  Here we use a language model again: *P(D).*

Translation model

$$\max_D P(D \mid E) = \max_D P(E \mid D) \cdot P(D)$$

- Data: parallel corpora
  (contain texts in two languages and their alignment)
- Most important corpus is Europarl: Data from the European Parliament, which translates all documents into all of the official EU languages.
- First step: need to align texts on a sentence-by-sentence basis.
- Task: estimate the probability of an English target sentence given the source sentence in German P(E|D).
- To avoid sparse data problems, we‘ll again have to use language models which make simplifying assumptions (like n-gram models of PCFGs).

Translation model: first naive attempt

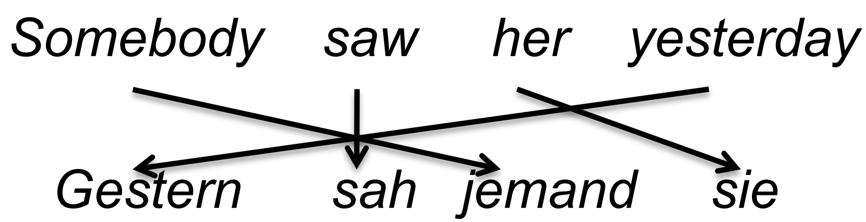
- We assume word-to-word translation: translation probability t(d|e) for German-English word pairs is independent of context. Then we get for sentence length n:

$$P(E \mid D) \approx \prod_{i=1}^{n} t(e_i \mid d_i)$$

- Beispiel:

$$\begin{array}{ccc} \textit{Somebody} & \textit{saw} & \textit{her} \\ \downarrow 0.8 & \downarrow 0.3 & \downarrow 0.5 \\ \textit{Jemand} & \textit{sah} & \textit{sie} \end{array}$$

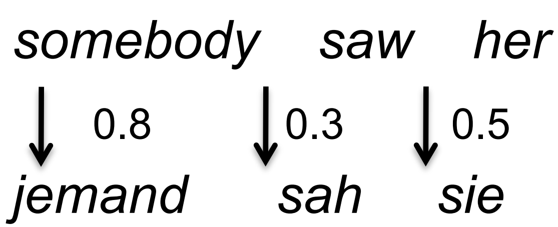
Translation model: word alignment

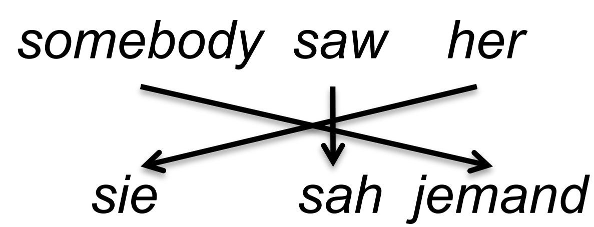


- Second attempt: Same equation, but i doesn‘t stand for word positions but for alignment pairs (in our example: (1,3), (2,2), (3,4), (4,1)).

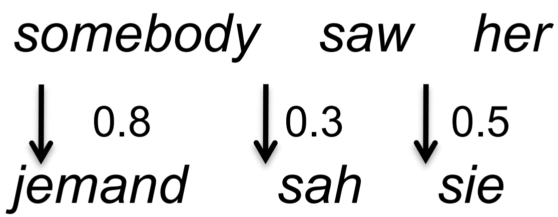
$$P(D \mid E) \approx \prod_{i=1}^{n} t(d_i \mid e_i)$$

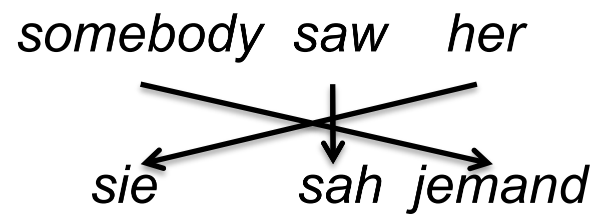
Problem: German has flexible word order.





Translation model: introduce reordering costs





$$P(E \mid D) \approx \prod_{i=1}^{n} t(e_i \mid d_i) * d(d_i, e_i)$$

$$d(d_i, e_i) = \frac{1}{2^{|d_i - e_i|}}$$

Challenges for word-by-word translation models

- Somebody saw her yesterday
- Gestern sah jemand sie

- Somebody saw her the day before yesterday
- Jemand sah sie vorgestern

- I guess somebody saw her
- Ich vermute, dass jemand sie sah

- I guess someboy saw her
- Ich vermute, dass jemand sie gesehen hat

- Somebody saw her
- Jemand hat sie gesehen

Translation model and language model

$$\max_D P(D \mid E) = \max_D P(E \mid D) \cdot P(D)$$

We can determine how „good“ a translation is by observing:

- How well it reflects the input.
  This is approximated by the translation model *P(E|D).*
- The fluency and correctness of the sentence in the target language.
  Here we use a language model again: *P(D).*

Language model for the target language

$$\max_D P(D \mid E) = \max_D P(E \mid D) \cdot P(D)$$

E: Bill worked at home yesterday

D1: Bill arbeitete zu Hause gestern

E: Bill worked at home yesterday

D2: Bill arbeitete gestern zu Hause

$$P(E \mid D1) > P(E \mid D2) \qquad P(E \mid D1) * P(D1) < P(E \mid D2) * P(D2)$$

- We use language models (such as bigrams) to deal with sparsity.

  $P(w_n|w_1w_2\ldots w_{n-1}) \approx P(w_n|w_{n-1})$

  $P(w_1w_2 \ldots w_n) \approx P(w_1)*P(w_2|w_1)*P(w_3|w_2)* \ldots P(w_n|w_{n-1})$

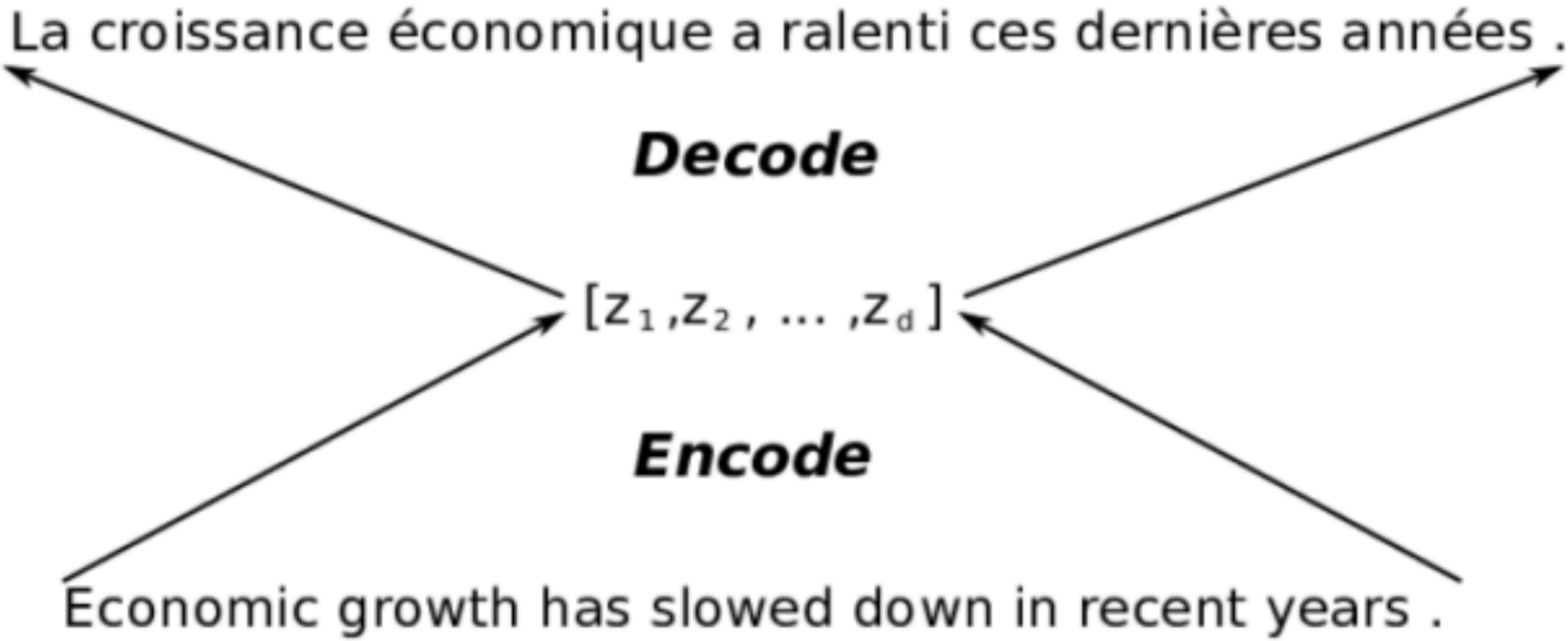
Summary Statistical MT

- Result quality comparable or better than best rule-based systems. (Depends on availability of parallel data.)
- Easier and faster to train on new languages or domains, compared to rule-based systems.
- Very good results if parallel corpora contain highly similar text to target application, because large patterns can be learned.

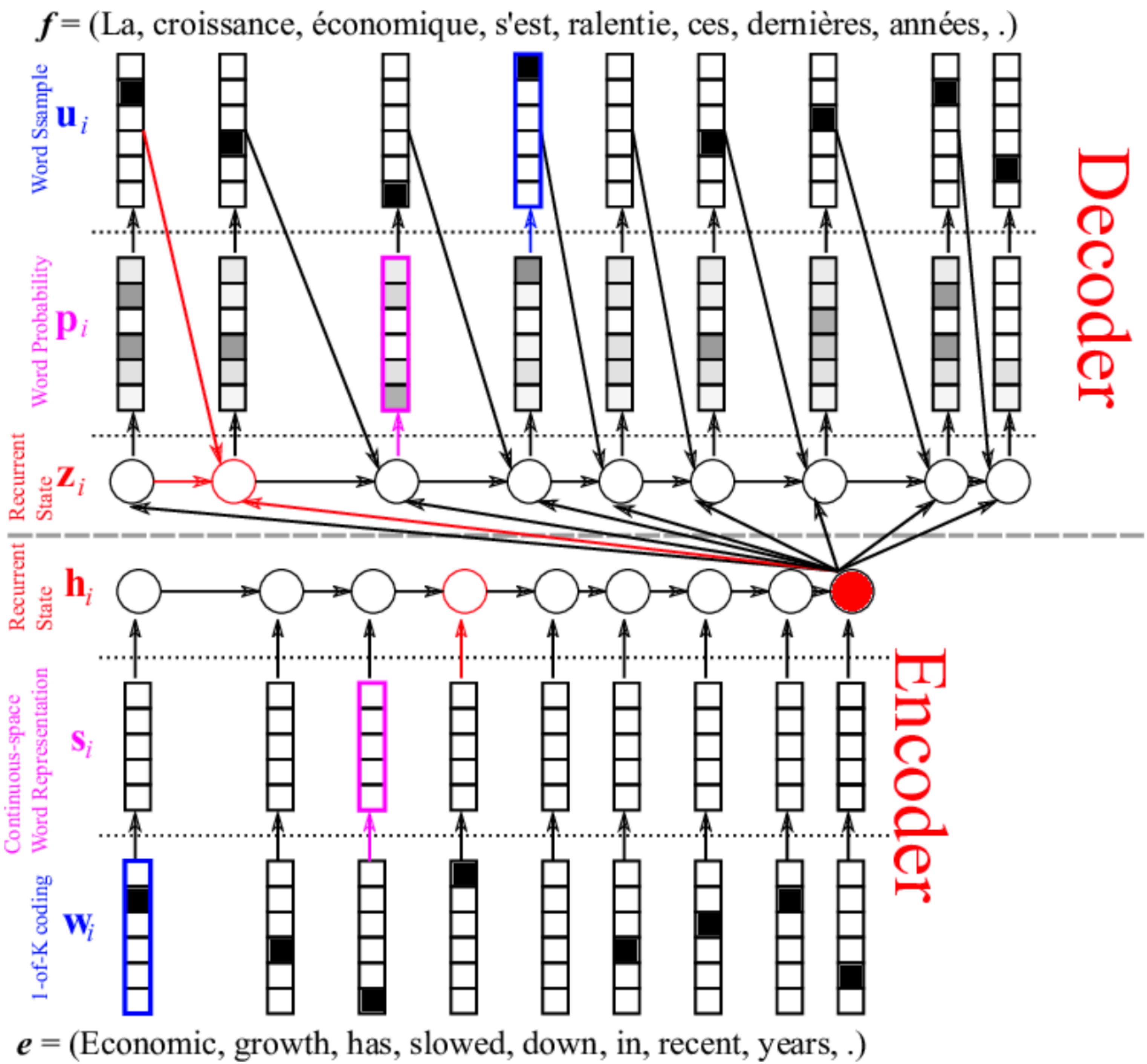
Contents for today

- Why is machine translation (MT) difficult?
- Approaches to MT
  - Knowledge-based (rule-based) MT
  - Statistical MT
  - **Neural MT**
- Evaluation

Seq2Seq Neural Net for MT

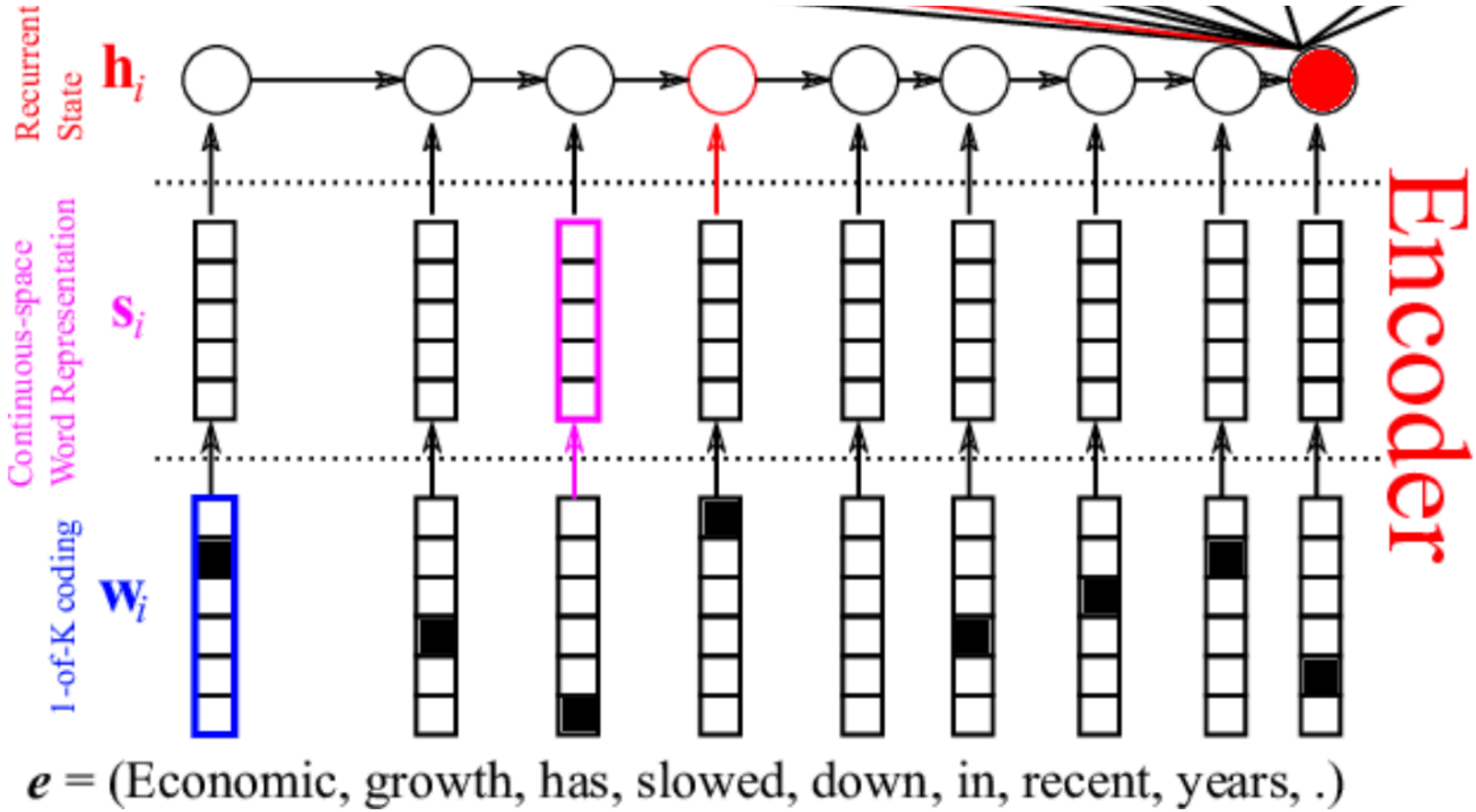


For a good explanation of the material in these slides, see
Devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-1/
Devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-2/
Devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-3/



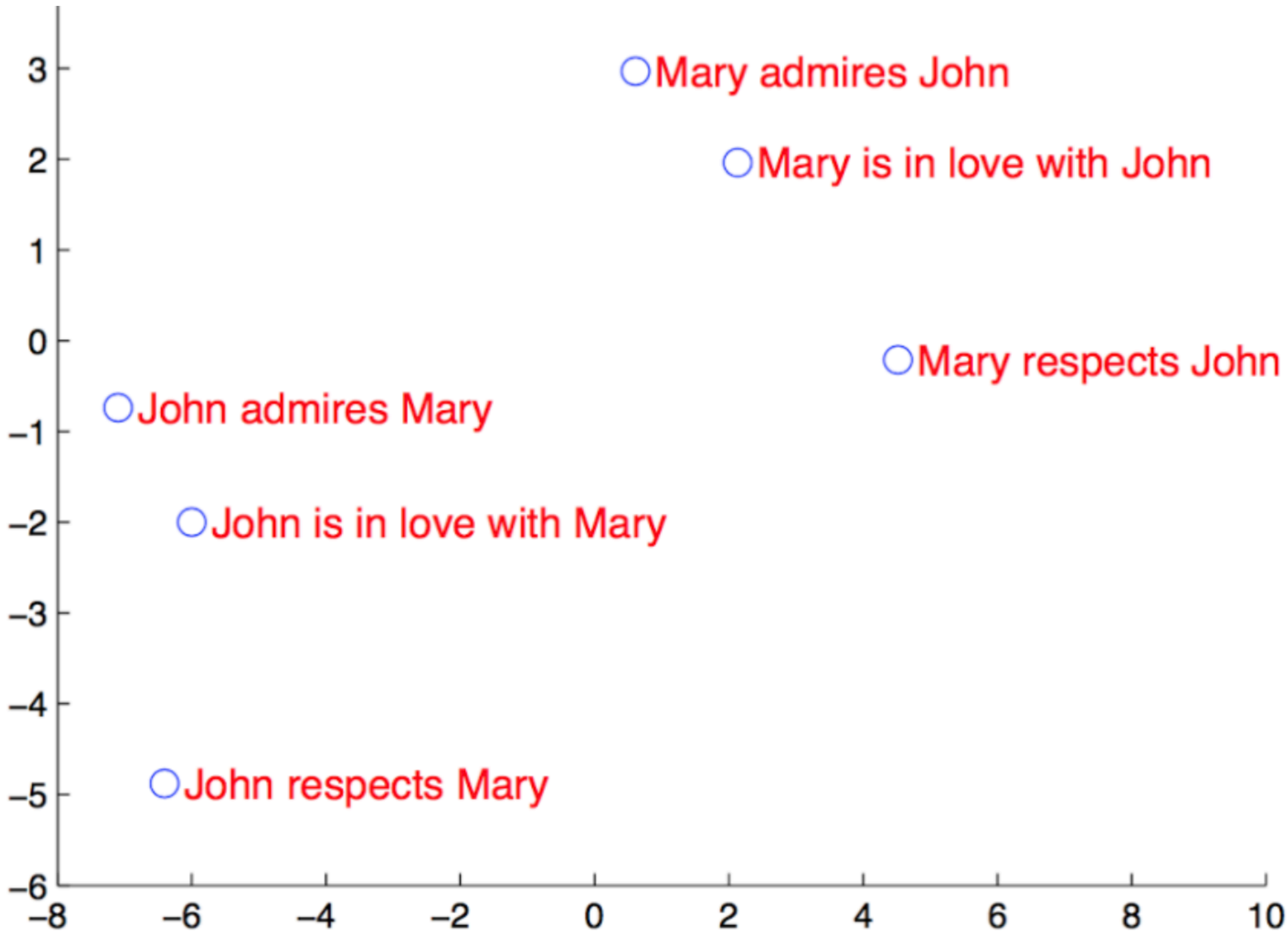
f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
u_i
Word Probability
p_i
Recurrent State
z_i
Decoder
Recurrent State
h_i
Continuous-space Word Representation
s_i
1-of-K coding
w_i
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

The encoder

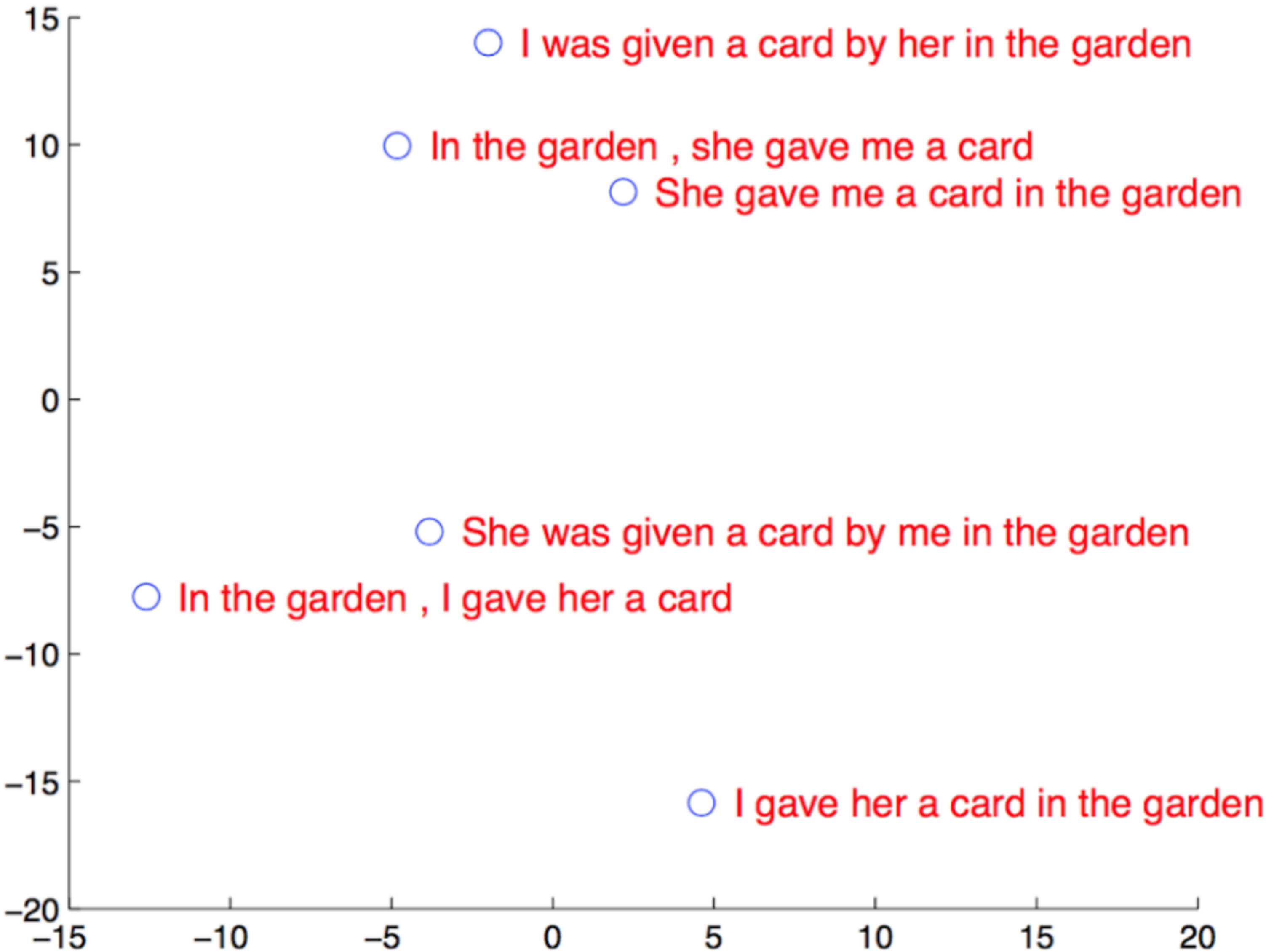


Recurrent State
$\mathbf{h}_i$
Continuous-space Word Representation
$\mathbf{s}_i$
1-of-K coding
$\mathbf{w}_i$
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

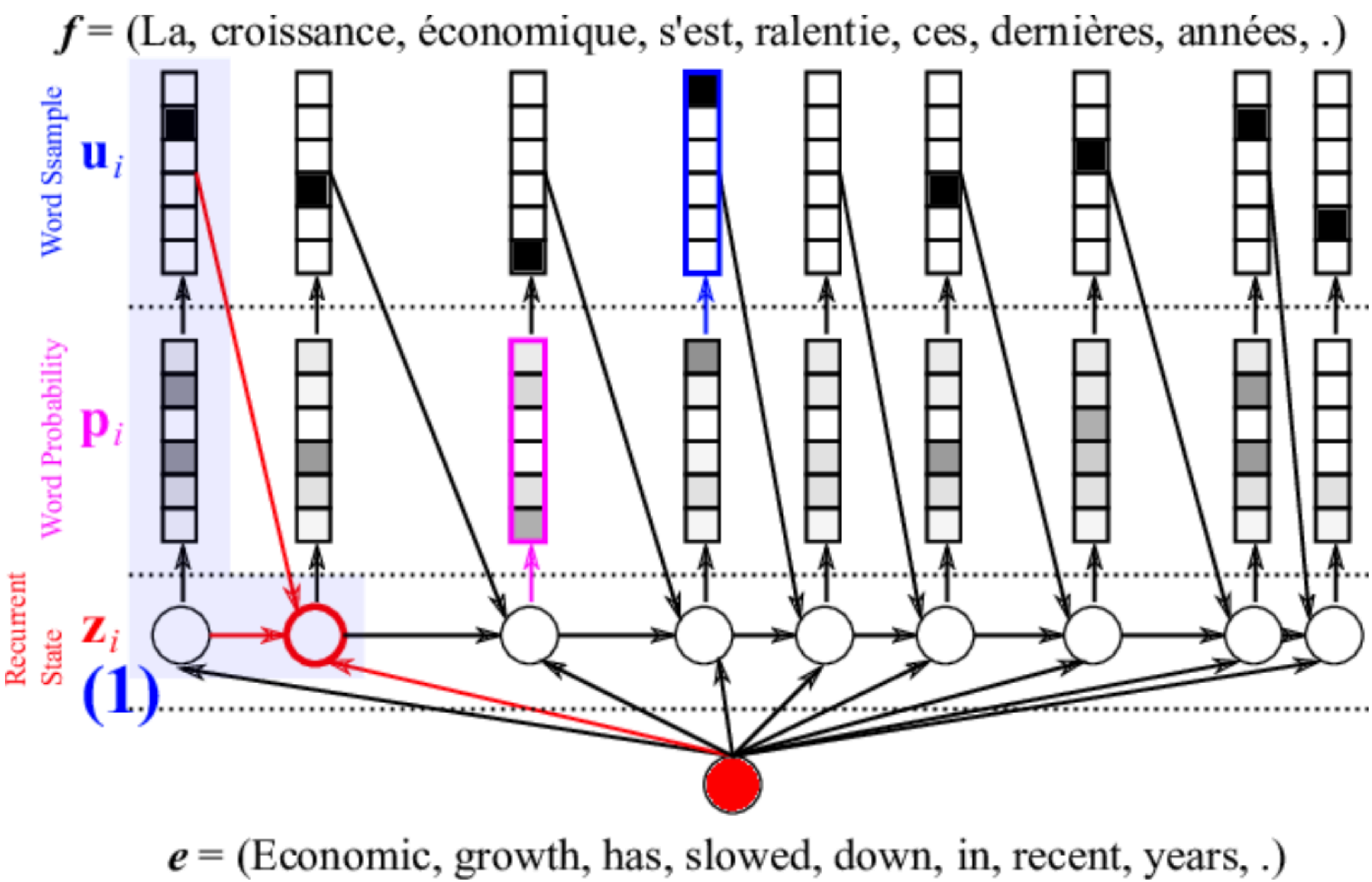
Example of what is represented after encoding



Example of what is represented after encoding



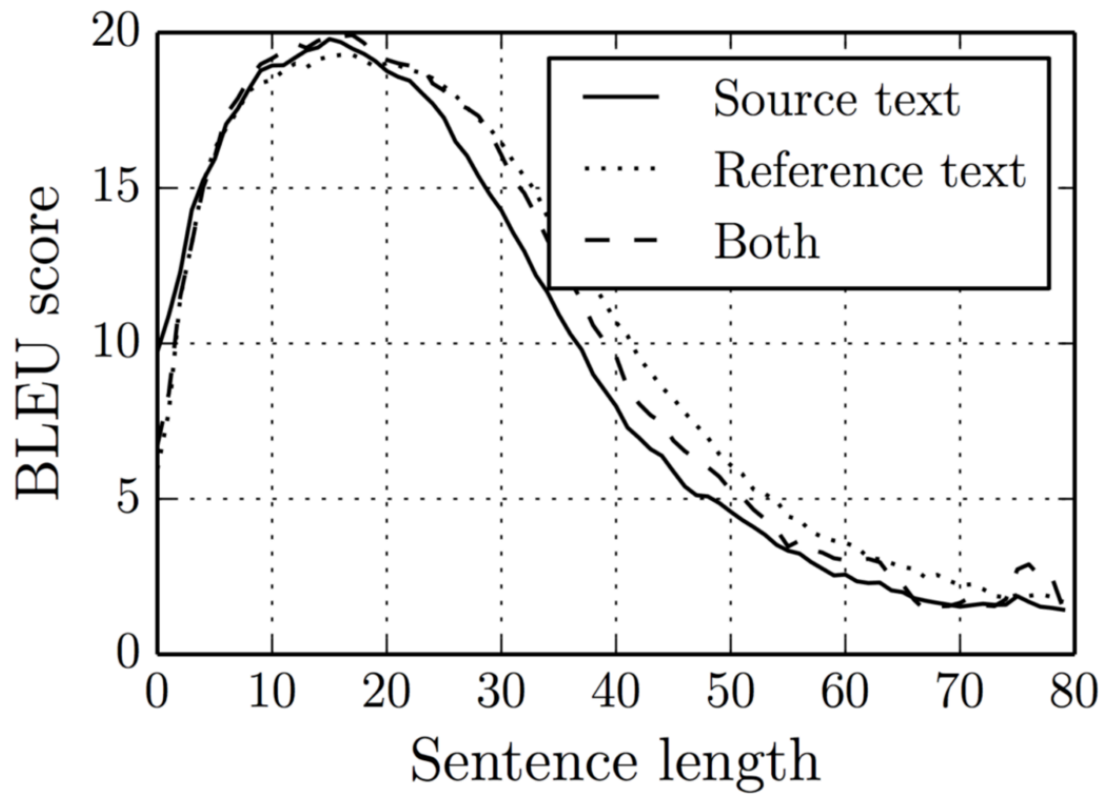
The decoder



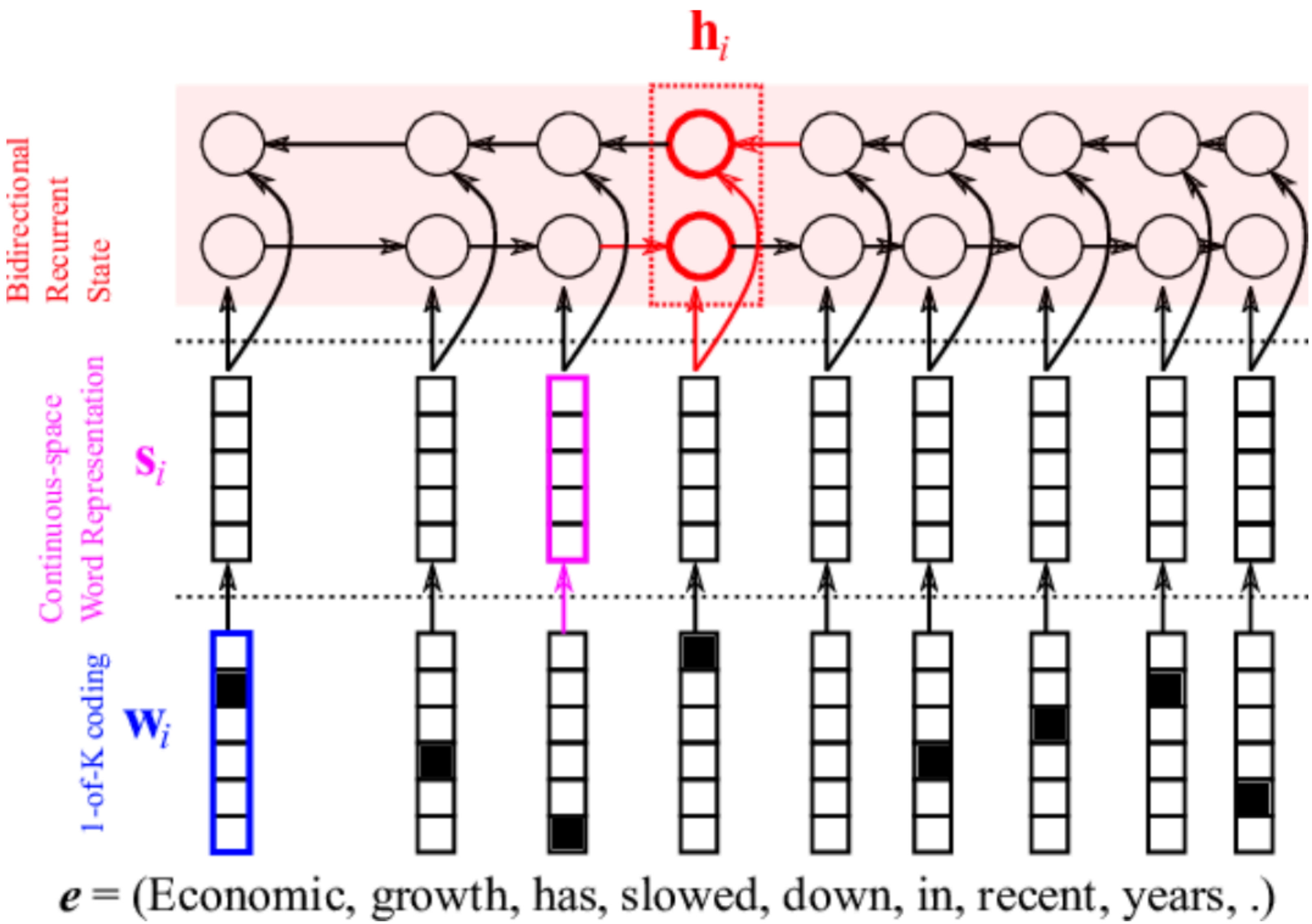
f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
$\mathbf{u}_i$
Word Probability
$\mathbf{p}_i$
Recurrent State
$\mathbf{z}_i$
(1)
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Challenge: encoding long sentences

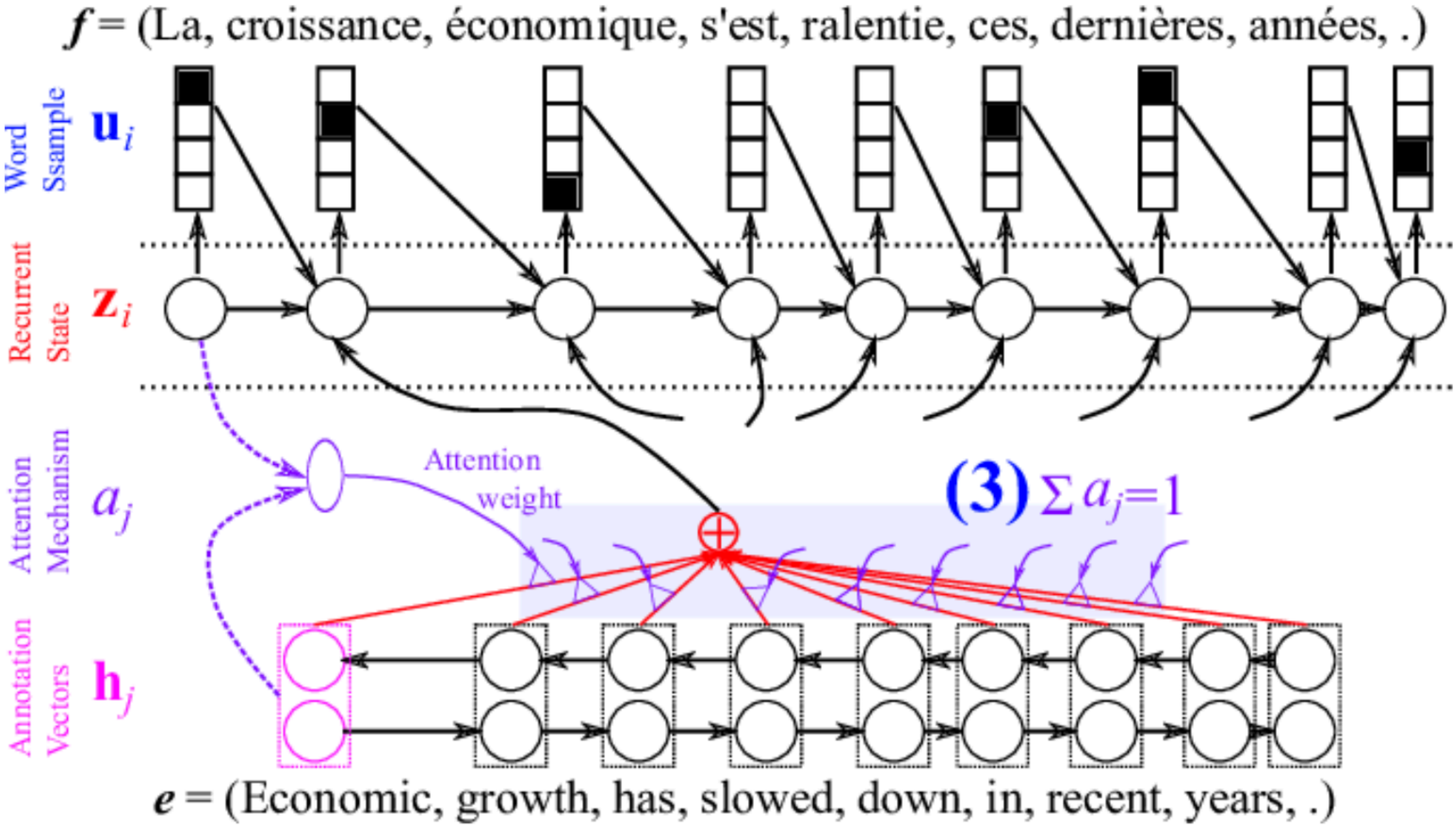
Sentences can be very long –
it doesn't work very well to try to store a very long sentence in a fixed sized vector



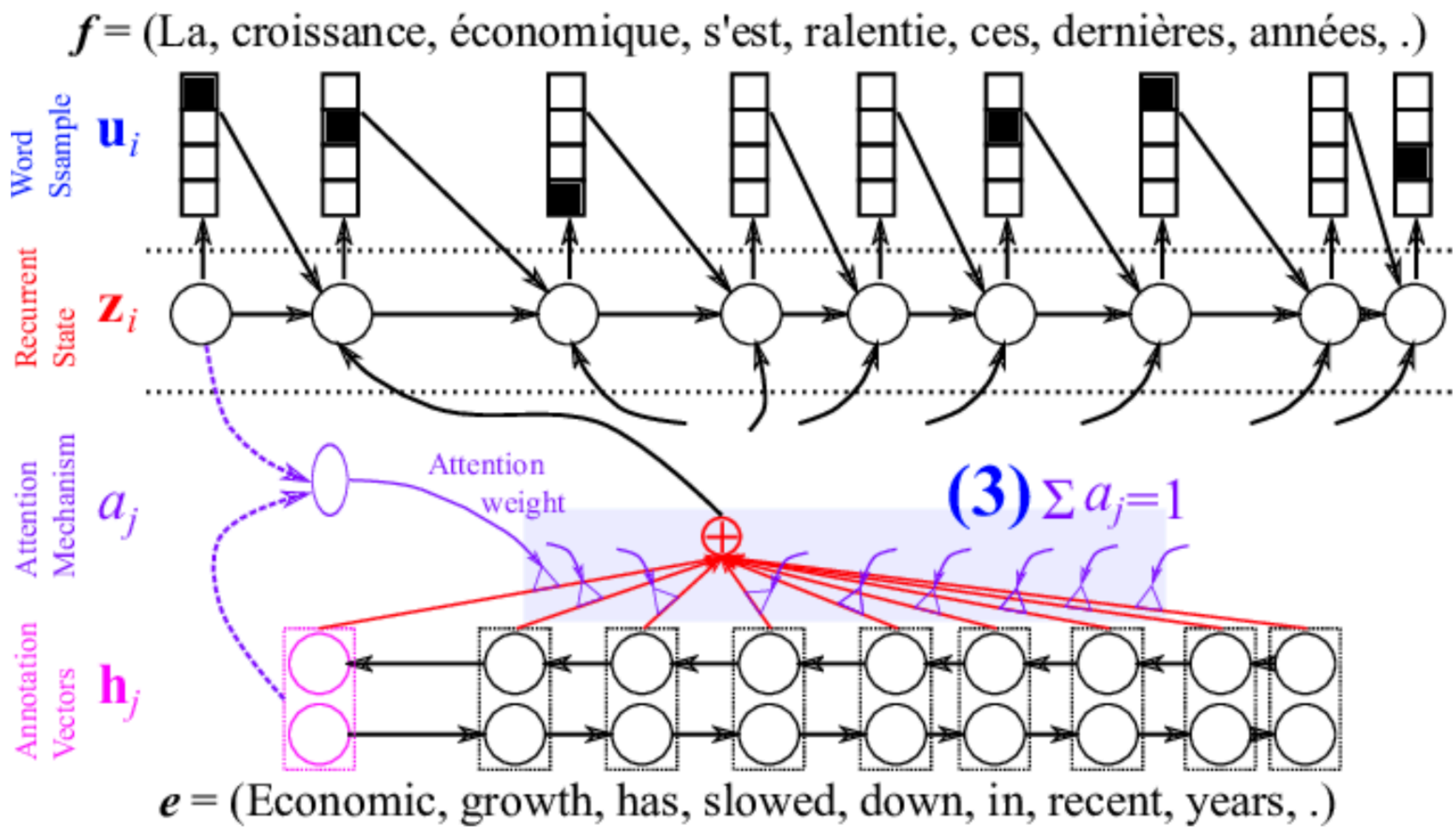
Alleviate problem by encoding from both sides



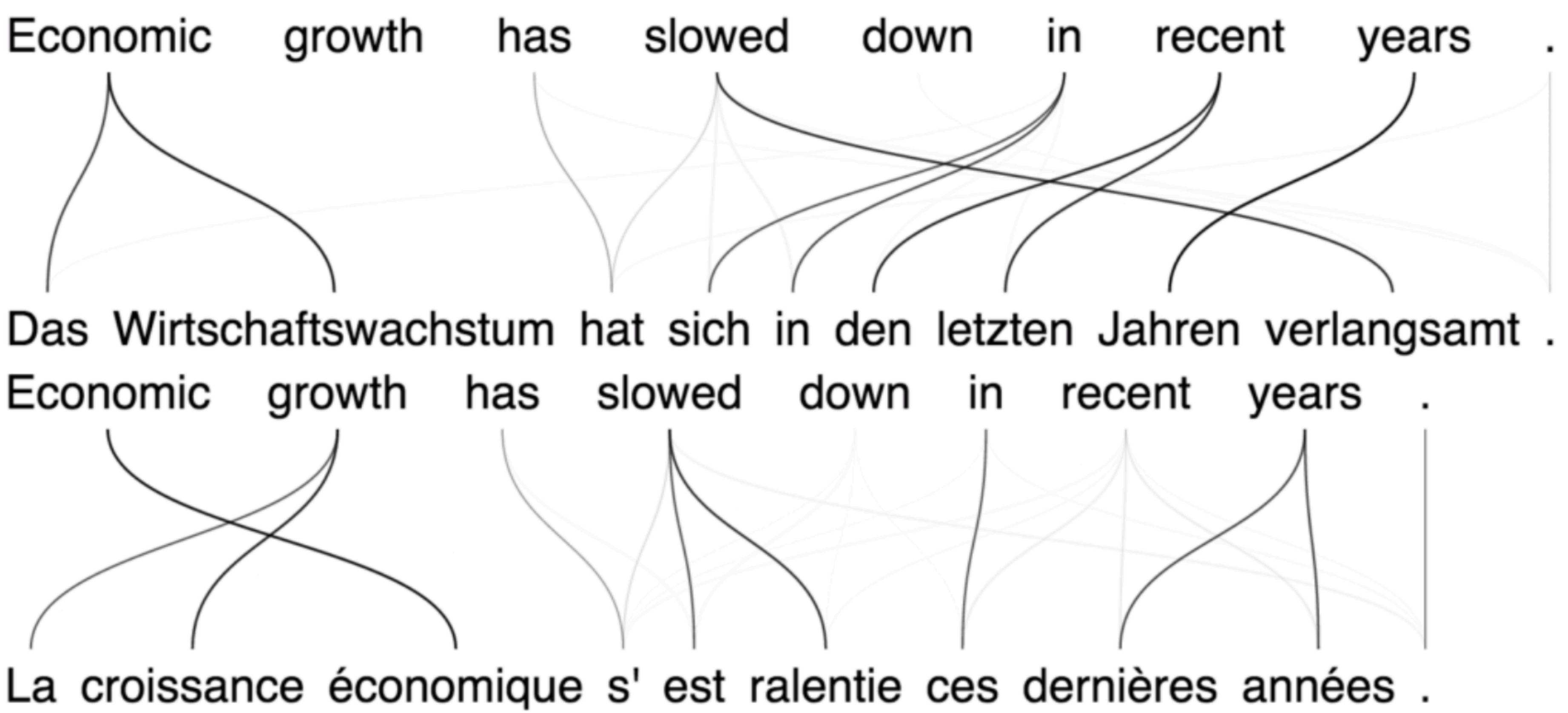
Add an attention mechanism



- For each position in the decoder, we learn where to look in the bidirectional representation from the encoder.
- Attention weights depend on what we have already said in decoder, as well as input word representations.
- Attention weights are re-weighted to achieve a probability distribution for each position.



Attention allows to focus on relevant parts of source text.



So, are computers able to translate?

Answer depends on

- Text type:

  Poetry vs. Legal documents vs. newspaper

- Expectations of the user:

  Precise information or rough idea of contents

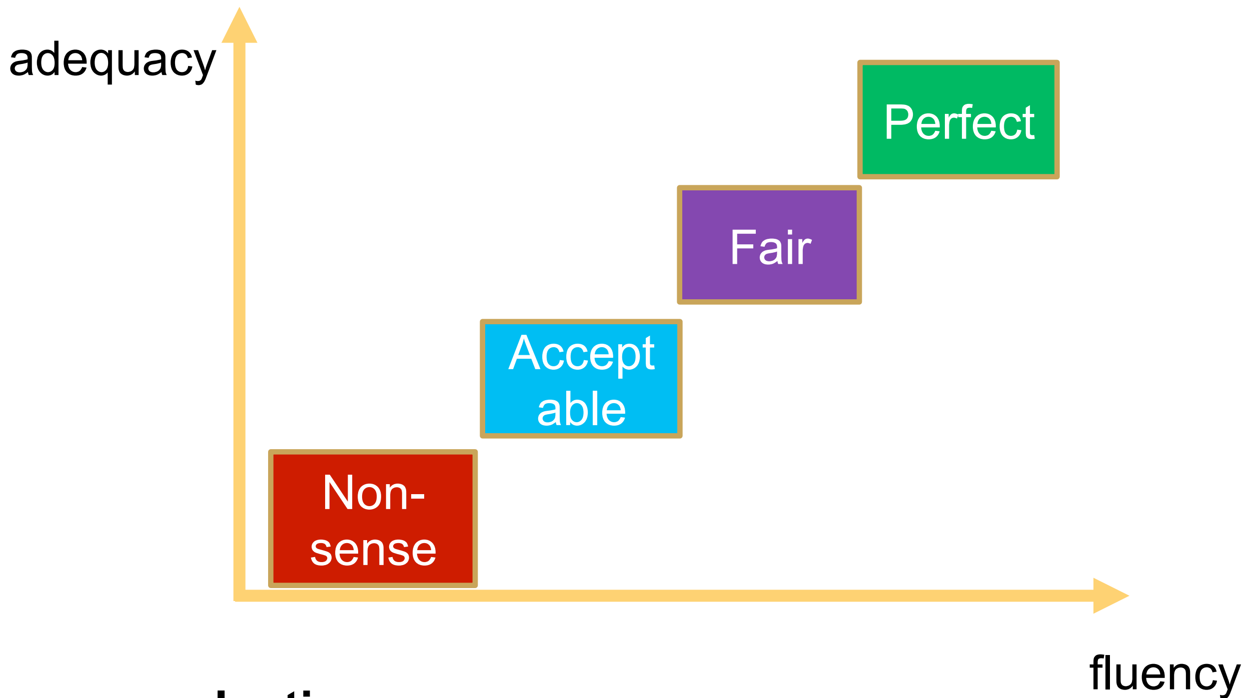
- „Leidensdruck“ des Nutzers:

  MT English → German vs. MT Chinese → German

Contents for today

- Why is machine translation (MT) difficult?
- Approaches to MT
  - Knowledge-based (rule-based) MT
  - Statistical MT
  - Neural MT
- **Evaluation**

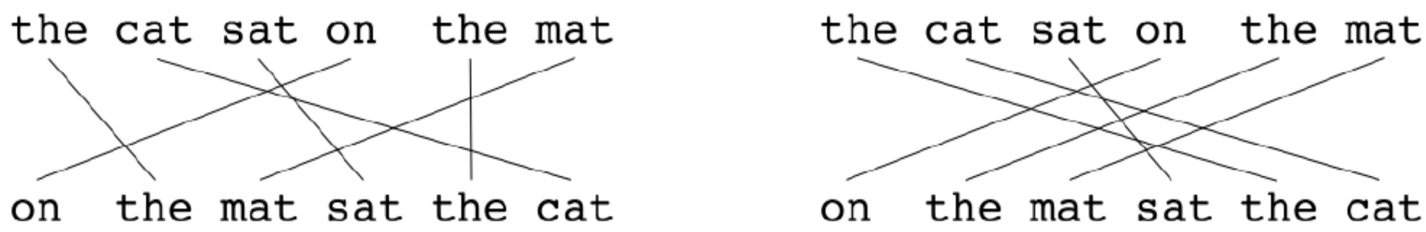
Expectation from a Good Evaluation Metric



- **Scale for human evaluation**
  - Perfect: No problem in both information and grammar
  - Fair: Easy to understand with some un-important information missing / flawed grammar
  - Acceptable: Broken but understandable with effort
  - Nonsense: important information has been realized incorrectly

Evaluation for MT / Natural Language Generation more generally

- Automatic metrics
  - Measure similarity with human generated texts
  - Word-over-lap based metrics, such as BLEU, METEOR, etc.



- Human Evaluation
  - Intrinsic: Fluency, Informativeness, Overall Quality
  - Extrinsic: Contribution to task success

Overlap Based Metrics

BLEU

- **B**i**L**ingual **E**valuation **U**nderstudy.
- Traditionally used for machine translation.
  - Ubiquitous and standard evaluation metric
  - 60% NLG works between 2012-2015 used BLEU
- Automatic evaluation technique:
  - **Goal**: ***The closer machine translation is to a professional human translation, the better it is*.**
- Precision based metric.
  - *How many results returned were correct?*
- Precision for NLG:
  - *How many words returned were correct?*

[Papineni et al., 2002]

BLEU evaluation

- **Candidate (Machine)**: *It is a guide to action which ensures that the military always obeys the commands of the party.*

- **References (Human)**:
  1. It is a guide to action that ensures that the military will forever heed Party commands.
  2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
  3. It is the practical guide for the army always to heed the directions of the party.

- Precision = $$\frac{\text{Total \#overlapping words}}{\text{Total \#words in candidate summary}} = \frac{17}{18}$$

[Papineni et al., 2002]

Problems with BLEU

- **Candidate**: the the the the the the the.

- **References**:
  1. The cat is on the mat.
  2. There is a cat on the mat.

- Unigram Precision = 7/7 = 1. Incorrect.
- Modified Unigram Precision = 2/7. (based on count clipping)
- Maximum reference count (‘the’) = 2
- Modified 1-gram precision → **Modified n-gram precision**.

[Papineni et al., 2002]

Modified n-gram precision

- **Candidate (Machine)**: *It is a guide to action which ensures that the military always obeys the commands of the party.*

- List all possible n-grams. (Example bigram : It is)
- N-gram Precision = $\dfrac{\text{Total \#overlapping n-grams}}{\text{Total \#n-grams in candidate summary}}$
- Modified N-gram Precision : ***Produced by clipping the counts for each n-gram to maximum occurrences in a single reference.***

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

[Papineni et al., 2002]

Brevity Penalty

- Candidate sentences longer than all references are already penalized by modified n-gram precision.

- Another multiplicative factor introduced.

- **Objective**: To ensure the candidate length matches one of the reference length.
  - If lengths equal, then BP = 1.
  - Otherwise, BP < 1.

[Papineni et al., 2002]

Final BLEU score

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

- BP → Brevity penalty.
- p_n → Modified n-gram precision.
- Number $N = 4$
- Weights $w_n = 1/N$

[Papineni et al., 2002]

Evaluation of data-to–text NLG: More BLUEs for BLEU

- **Intrinsically Meaningless** (Ananthakrishnan et al, 2009)
  - Not meaningful in itself: What does a BLEU score of 69.9 mean?
  - Only for comparison between two or more automatic systems
- **Admits too much “combinatorial” variation**
  - Many possible variations of syntactically and semantically incorrect variations of hypothesis output
  - Reordering within N-gram mismatch may not alter the BLEU scores
- **Admits too little “linguistic” variation**
  - Languages allow variety in choice of vocabulary and syntax
  - Not always possible to keep all possible variations as references
  - Multiple references do not help capture variations much (Doddington, 2002; Turian et al, 2003)
- **Variants of BLEU**: cBLEU (Mei et al, 2016), GLEU (Mutton et al, 2007), Q-BLEU (Nema et al, 2018), take input (source) into account

ROUGE

- Recall-Oriented Understudy for Gisting Evaluation.
- Recall based metric for NLP:
  - *How many correct words were returned?*

- **Candidate**: the cat was found under the bed.
- **Reference**: the cat was under the bed.
- Recall = $\dfrac{\text{Total \#overlapping words}}{\text{Total \#words in reference summary}} = \dfrac{6}{6}$

- ROUGE metric:

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

[Lin 2004]

Problems with overlap based metrics

- References needed
- Assumes output space to be confined to a set of reference given
- Often penalizes paraphrases at syntactic and deep semantic levels
- Task agnostic
  - Cannot reward task-specific correct generation
- Relativistic evaluation
  - Intrinsically don’t mean anything (*what does 50 BLEU mean?)*

BLEU not perfect for evaluation…..

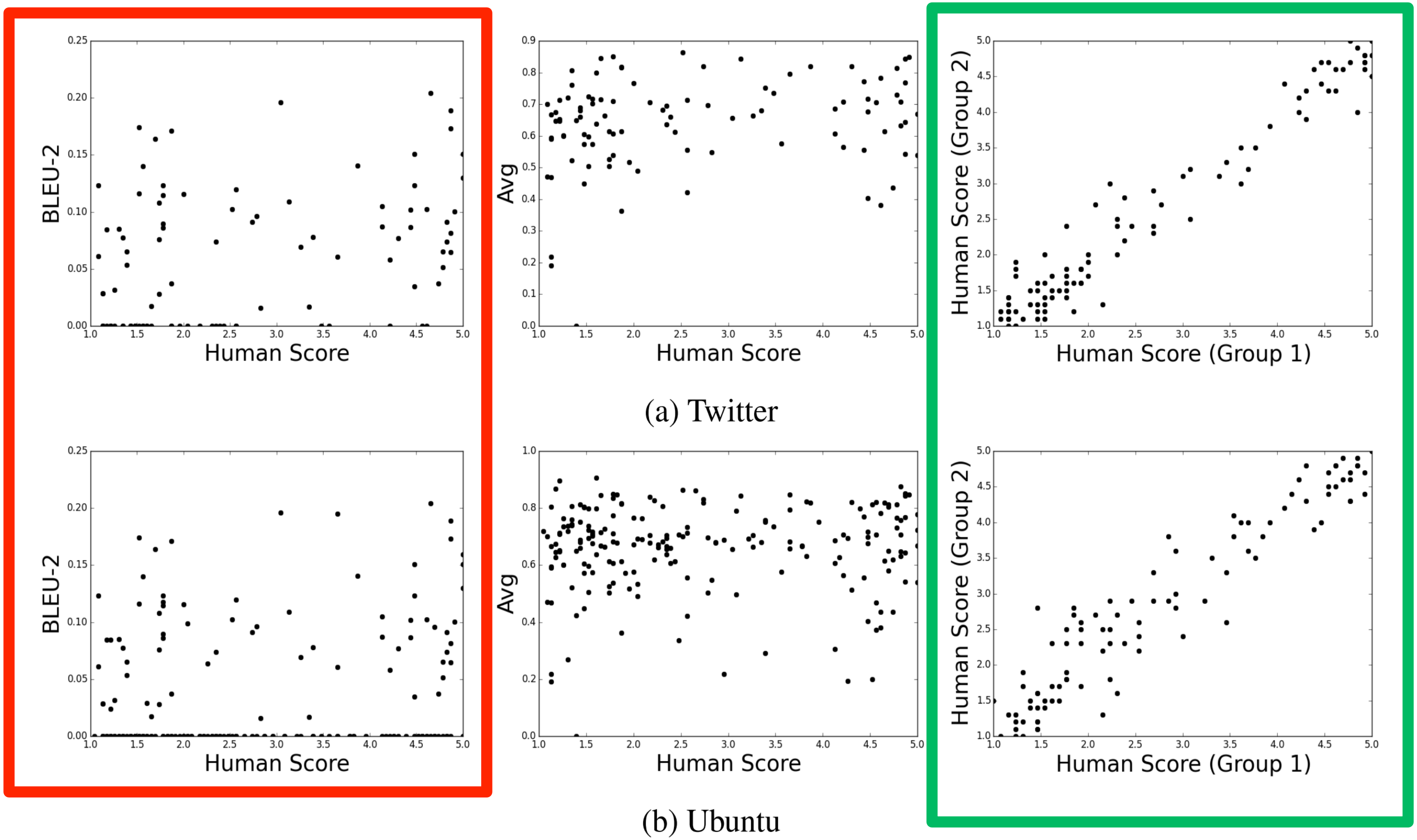


Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

[Liu et al., 2016]

ROUGE comes at a cost….

- **[Paulus et al., 2017] used Reinforcement Learning (RL) to directly optimize for ROUGE-L**
  - Instead of the usual cross-entropy loss.
  - ROUGE-L is not differentiable, hence need RL-kind of framework.
- **Observation**:
  - Outputs obtained with **higher** ROUGE-L scores, but **lower** human scores for relevance and readability.

Model	ROUGE-1	ROUGE-2	ROUGE-L
ML, no intra-attention, no trigram avoidance	42.85	26.22	39.09
ML, no intra-attention	44.26	27.43	40.41
ML, with intra-attention	43.86	27.10	40.11
RL, no intra-attention	**47.22**	30.51	**43.27**
ML+RL, no intra-attention	47.03	**30.72**	43.10

Model	Readability	Relevance	Perplexity
ML	6.76	7.14	**84.46**
RL	4.18	6.32	16417.68
ML+RL	**7.04**	**7.45**	121.07

berg

Slide credit: CS224n, Stanford
[Paulus et al., 2017]

Summary

- No Automatic metrics to adequately capture overall quality of generated text (w.r.t human judgement).

- Though more focused automatic metrics can be defined to capture particular aspects:
  - **Fluency** (compute probability w.r.t. well-trained Language Model).
  - **Correct Style** (probability w.r.t. LM trained on target corpus – still not perfect)
  - **Diversity** (rare word usage, uniqueness of n-grams, entropy-based measures)
  - **Relevance to input** (semantic similarity measures – may not be good enough)
  - Simple measurable aspects like **length** and **repetition**
  - **Task-specific metrics**, e.g. compression rate for summarization

Human Evaluation

Human judgement scores typically considered in NLG

- **Fluency:** How grammatically correct is the output sentence?

"Ah, go boil yer heads, both of yeh. Harry—yer a wizard."

- **Adequacy:** To what extent has information in the input been preserved in the output ?

INPUT: <Einstein, birthplace, Ulm> | OUTPUT: Einstein was born in Florence

- **Coherence:** How coherent is the output paragraph?

The most important part of an essay is the thesis statement. Essays can be written on various topics from domains such as politics, sports, current affairs etc. I like to write about Football because it is the most popular team sport played at international level.

- **Readability:** How hard is the output to comprehend?

A neutron walks into a bar and asks how much for a drink. The bartender replies "for you no charge."

Problems with human evaluation

- **Can be slow and expensive**
- **Can be unreliable:**
  - Humans are (1) inconsistent, (2) sometimes illogical, (3) can lose concentration, (4) misinterpret the input, (5) cannot always explain why they feel the way they do.
- **Can be subjective** (vary from person to person)
- **Judgements can be affected by different expectations**
  - "the chatbot was very engaging because it always wrote back"

Better automatic evaluation metrics are needed!!!!