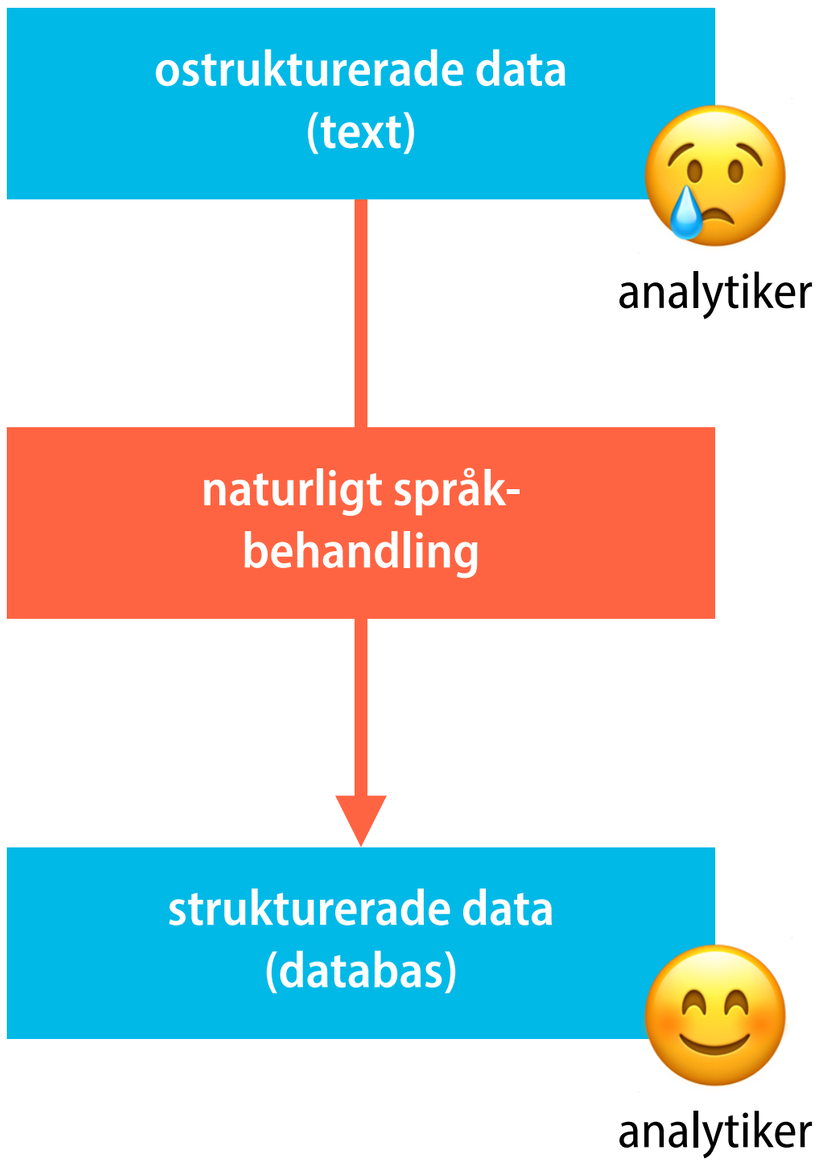
AI för naturligt språk

Automatisk språkförståelse

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Kunskapsglappet



Informationsextraktion med mallar

As of 15 Mar 2002, Hawaii state health officials reported one additional recent case of dengue fever and 6 cases that occurred last year but were not confirmed by laboratory testing until 2002.

Källa: Grishman et al. (2002)

Attribut	Värde
docno	ProMed.20020322.11
doc_date	2002.03.22
disease_name	dengue fever
norm_stime	2002.03.15
norm_etime	2002.03.15
victim_types	—
location	Hawaii

Namngivna entiteter

Hon är känd för sitt arbete inom naturligt språk-behandling och föddes i Huddersfield.

University of Cambridge, CC BY 2.0, via Wikimedia Commons

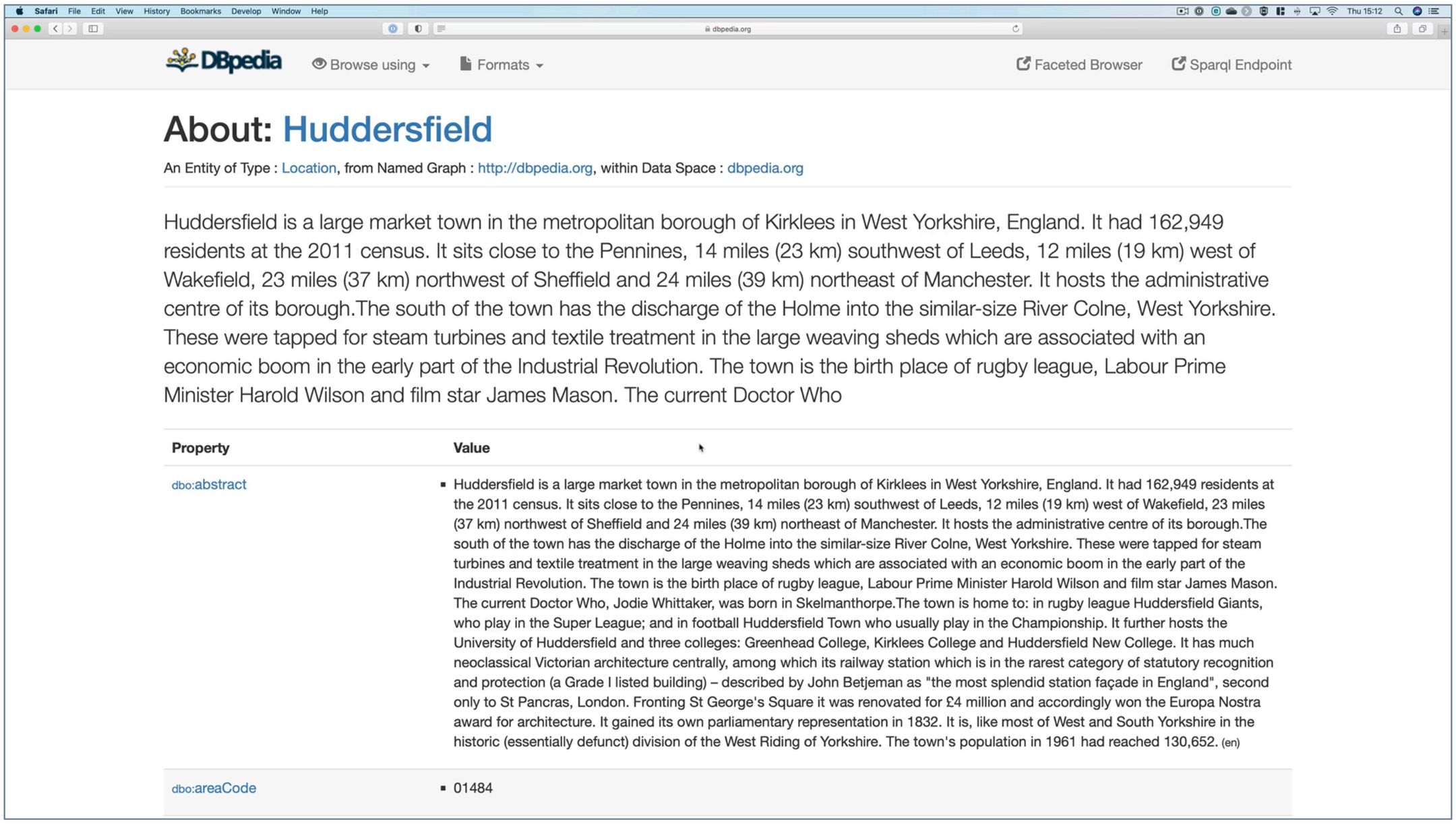
```
SELECT DISTINCT ?x WHERE {
  ?x dbo:knownFor dbr:Natural_language_processing.
  ?x dbo:birthPlace dbr:Huddersfield.
}
```

Semantiska relationer

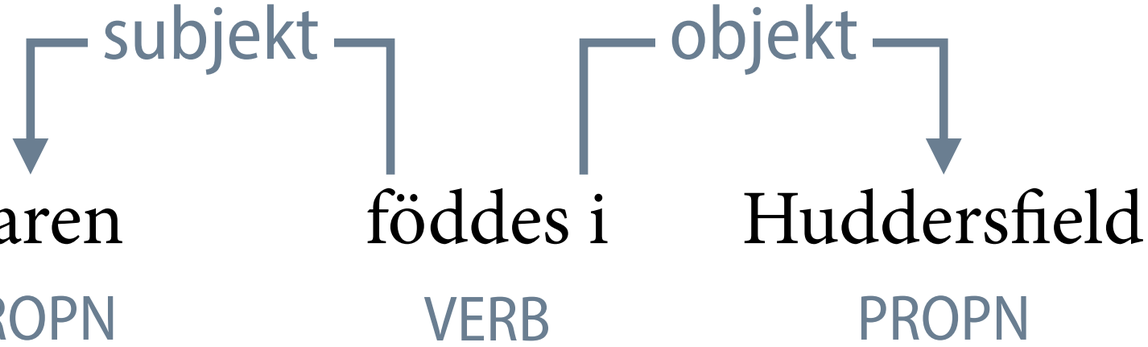
Hon är känd för sitt arbete inom naturligt språk-behandling och föddes i Huddersfield.

University of Cambridge, CC BY 2.0, via Wikimedia Commons

```
SELECT DISTINCT ?x WHERE {
  ?x dbo:knownFor dbr:Natural_language_processing.
  ?x dbo:birthPlace dbr:Huddersfield.
}
```


Safari File Edit View History Bookmarks Develop Window Help
Thu 15:12
dbpedia.org
DBpedia
Browse using
Formats
Faceted Browser
Sparql Endpoint
About: Huddersfield
An Entity of Type : Location, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org
Huddersfield is a large market town in the metropolitan borough of Kirklees in West Yorkshire, England. It had 162,949 residents at the 2011 census. It sits close to the Pennines, 14 miles (23 km) southwest of Leeds, 12 miles (19 km) west of Wakefield, 23 miles (37 km) northwest of Sheffield and 24 miles (39 km) northeast of Manchester. It hosts the administrative centre of its borough.The south of the town has the discharge of the Holme into the similar-size River Colne, West Yorkshire. These were tapped for steam turbines and textile treatment in the large weaving sheds which are associated with an economic boom in the early part of the Industrial Revolution. The town is the birth place of rugby league, Labour Prime Minister Harold Wilson and film star James Mason. The current Doctor Who
Property
Value
dbo:abstract
Huddersfield is a large market town in the metropolitan borough of Kirklees in West Yorkshire, England. It had 162,949 residents at the 2011 census. It sits close to the Pennines, 14 miles (23 km) southwest of Leeds, 12 miles (19 km) west of Wakefield, 23 miles (37 km) northwest of Sheffield and 24 miles (39 km) northeast of Manchester. It hosts the administrative centre of its borough.The south of the town has the discharge of the Holme into the similar-size River Colne, West Yorkshire. These were tapped for steam turbines and textile treatment in the large weaving sheds which are associated with an economic boom in the early part of the Industrial Revolution. The town is the birth place of rugby league, Labour Prime Minister Harold Wilson and film star James Mason. The current Doctor Who, Jodie Whittaker, was born in Skelmanthorpe.The town is home to: in rugby league Huddersfield Giants, who play in the Super League; and in football Huddersfield Town who usually play in the Championship. It further hosts the University of Huddersfield and three colleges: Greenhead College, Kirklees College and Huddersfield New College. It has much neoclassical Victorian architecture centrally, among which its railway station which is in the rarest category of statutory recognition and protection (a Grade I listed building) – described by John Betjeman as "the most splendid station façade in England", second only to St Pancras, London. Fronting St George's Square it was renovated for £4 million and accordingly won the Europa Nostra award for architecture. It gained its own parliamentary representation in 1832. It is, like most of West and South Yorkshire in the historic (essentially defunct) division of the West Riding of Yorkshire. The town's population in 1961 had reached 130,652. (en)
dbo:areaCode
01484

Taggning och parsning



```
dbr:Karen_Sparck_Jones dbo:birthPlace dbr:Huddersfield
```

AI för naturligt språk

Sekvenstaggning

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Textklassificering

etikett

sekvens av ord

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ \operatorname{score}(\boldsymbol{x}, y; \boldsymbol{\theta})$$

möjliga etiketter

modell-parametrar

exempel: polaritet

Sekvenstaggning

sekvens av etiketter

sekvens av ord

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, \operatorname{score}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

möjliga sekvenser av etiketter

modell-parametrar

exempel: ordklasser

Ordklasstaggning

Jag	bad	om	en	kort	bit
PRON	VERB	ADP	DET	ADJ	NOUN
NOUN	NOUN	SCONJ	NUM	ADV	VERB
		ADV	PRON	NOUN	
			NOUN		

Exempel från Joakim Nivre

Metoder för sekvenstaggning

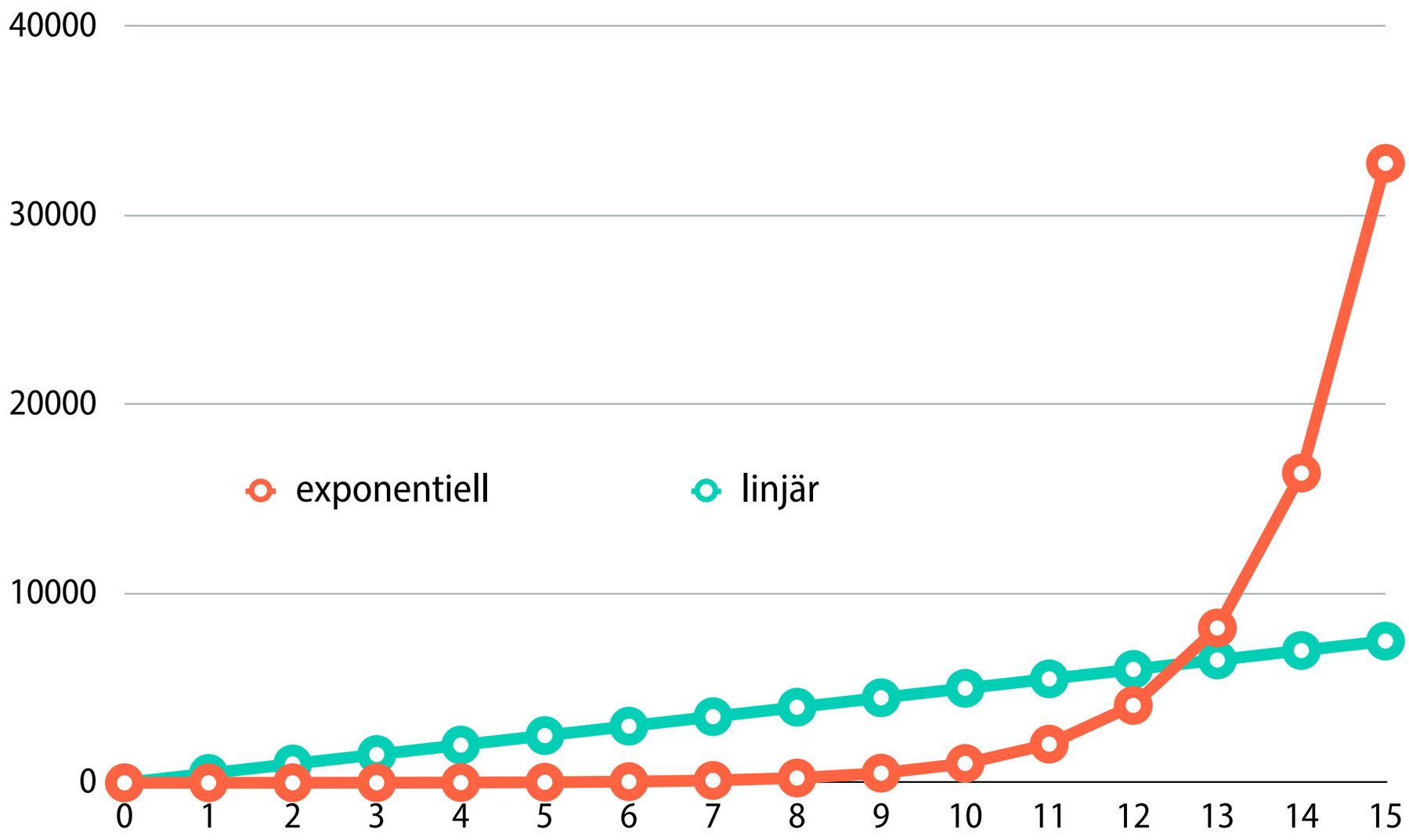
- **Lokal sökning**

  Formulera problemet som en sekvens av klassifikationsproblem: predicera etiketten för varje position i input-sekvensen.

- **Global sökning**

  Formulera problemet som ett optimeringsproblem: hitta den bästa möjliga etiketteringen av input-sekvensen.

Kombinatorisk explosion



AI för naturligt språk

Tillämpningar av sekvenstaggning 1

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Ordklasstaggning

Jag	bad	om	en	kort	bit
PRON	VERB	ADP	DET	ADJ	NOUN
NOUN	NOUN	SCONJ	NUM	ADV	VERB
		ADV	PRON	NOUN	
			NOUN		

Exempel från Joakim Nivre

Universella ordklasser

Källa: Universal Dependencies Project

Tagg	Kategori	Exempel
ADJ	adjektiv	*glad*
ADV	adverb	*inte*
INTJ	interjektion	*aj!*
NOUN	substantiv	*pudding*
PROPN	egennamn	*Mats*
VERB	verb	*kasta*

Tagg	Kategori	Exempel
ADP	adposition	*av*
AUX	hjälpverb	*har*
CCONJ	konjunktion	*och*
DET	determinerare	*denna*
NUM	grundtal	*tre*
PRON	pronomen	*hon*

plus PART, SCONJ, PUNCT, SYM, X

Textsegmentering

我	有	一 台	计 算 机	。
Jag	har	en	dator	.

Wang och Xu (2017)

Igenkänning av namngivna entiteter

ORG ORG

Denna alumna från Stanford University grundade utbildningsföretaget Coursera.

```
SELECT DISTINCT ?x WHERE {
  ?x dbo:almaMater dbr:Stanford_University.
  dbr:Coursera dbo:foundedBy ?x.
}
```

Aspektbaserad sentimentanalys

NEGATIV ASPEKT

Jag hatade deras fajitas,

men salladerna var jättegoda!

ASPEKT POSITIV

{fajitas: negativ, salladerna: positiv}

AI för naturligt språk

Tillämpningar av sekvenstaggning 2

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Igenkänning av namngivna entiteter som taggning

1	2	3	4	5	6	7	8	9
European	Union	rejects	German	call	to	boycott	British	lamb
B-ORG	**I-ORG**	**O**	**B-GPE**	**O**	**O**	**O**	**B-GPE**	**O**
början på en ORG	inuti en ORG			utanför en entitet			början på en GPE	

{(1, 2): ORG, (4, 4): GPE, (8, 8): GPE}

Textsegmentering som taggning

我	有	一 台	计 算 机	。
S	**S**	**B E**	**B M E**	**S**
single		beginning end	middle	

Wang och Xu (2017)

Utvärdering av sekvenstaggning

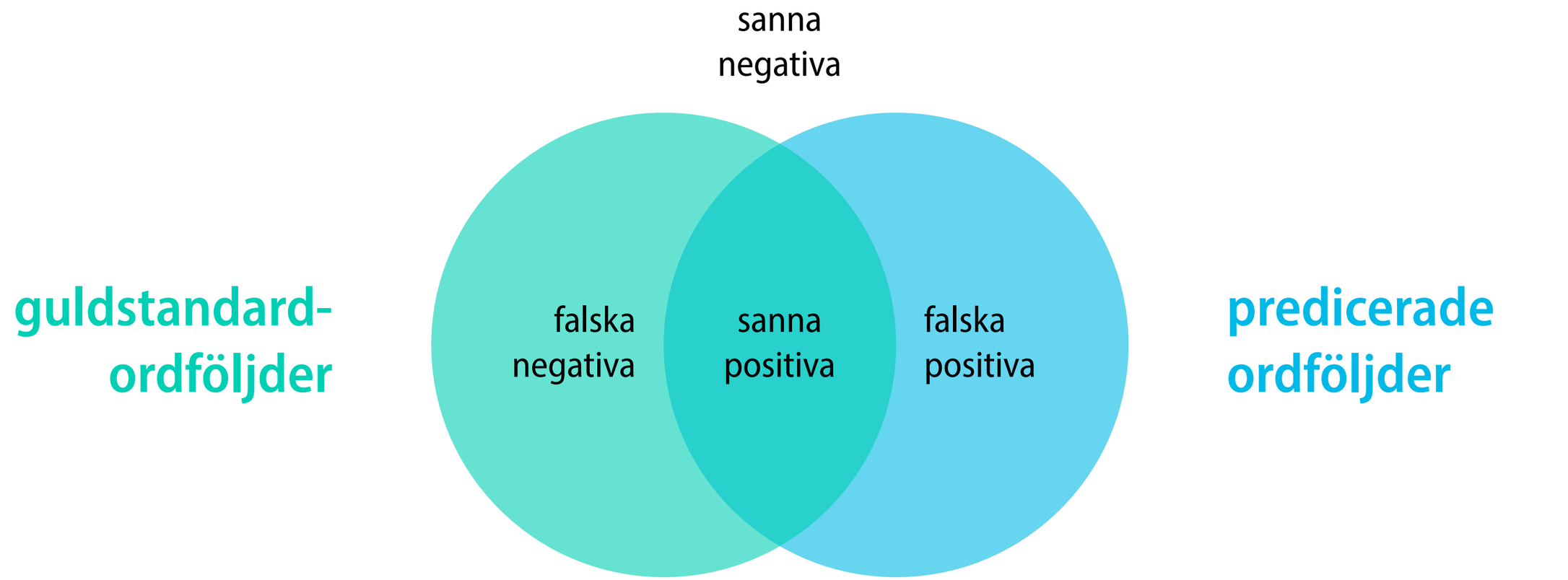
- Det vanligaste utvärderingsmåttet för ordklasstaggning är **korrekthet** (eng. *accuracy*) på ordnivå.

  andelen korrekt predicerade taggar

- De vanligaste utvärderingsmåtten för segmentering och namnigenkänning är **precision** och **täckning** (eng. *recall*).

  andelen korrekt predicerade/guldstandard-ordföljder

Precision och täckning för ordföljder



$$P = \frac{|\text{guld} \cap \text{predicerad}|}{|\text{predicerad}|}$$

$$R = \frac{|\text{guld} \cap \text{predicerad}|}{|\text{guld}|}$$

AI för naturligt språk

Sekvenstaggning med lokal sökning

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Metoder för sekvenspredicering

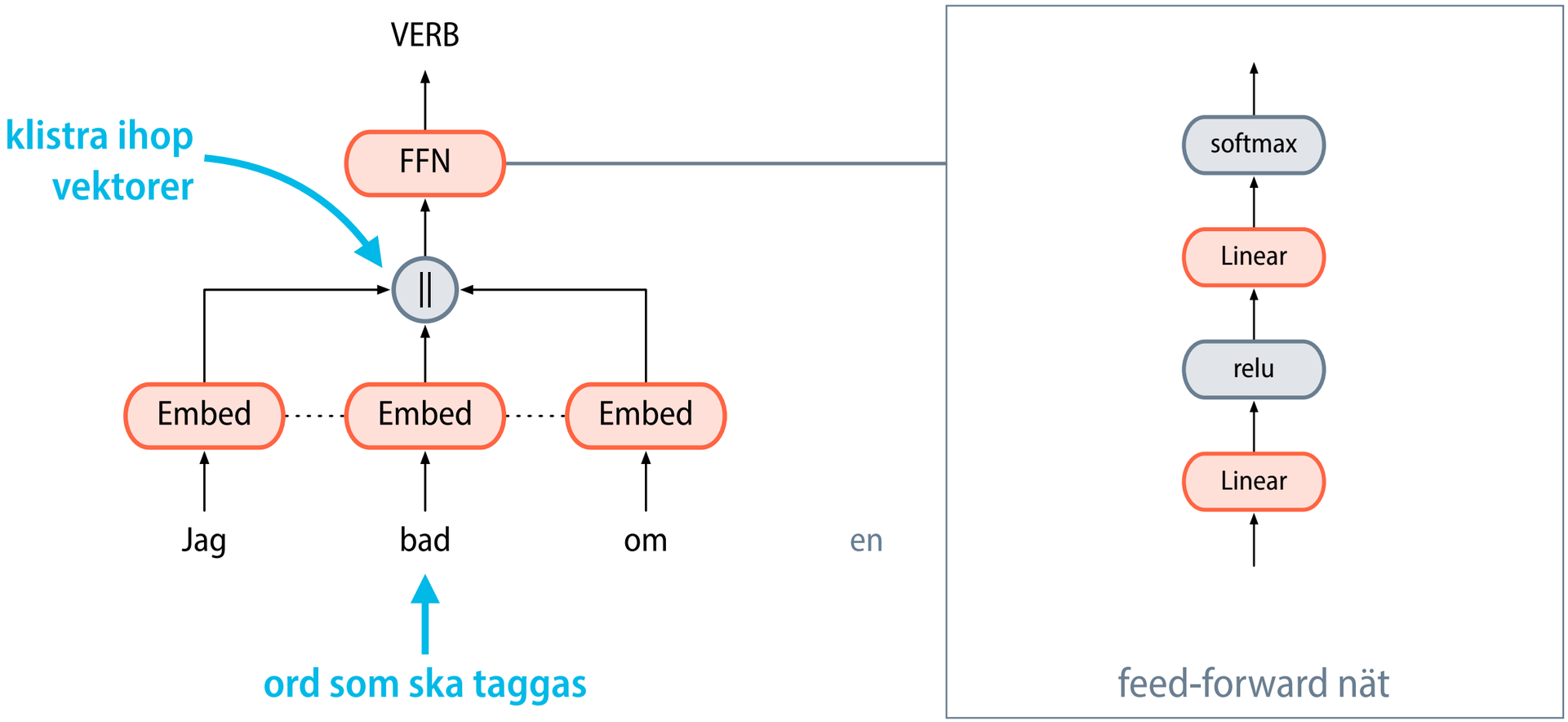
- **Lokal sökning**

  Formulera problemet som en sekvens av klassifikationsproblem: predicera etiketten för varje position i input-sekvensen.

- **Global sökning**

  Formulera problemet som ett optimeringsproblem: hitta den bästa möjliga etiketteringen av input-sekvensen.

Ordklasstaggning med fönstermodellen

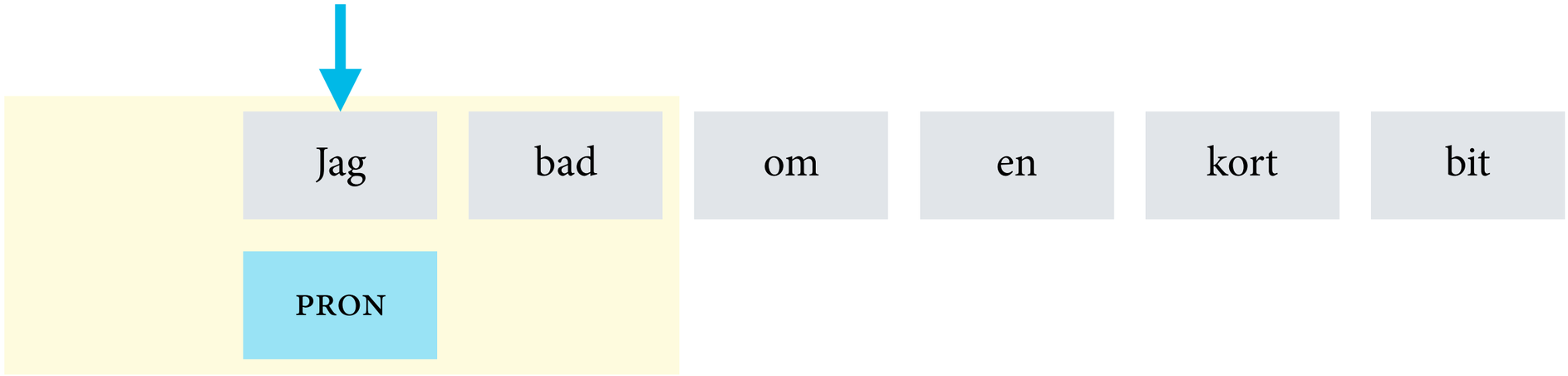


Taggarna är inte oberoende av varandra

Jag	bad	om	en	kort	bit	
PRON	VERB	ADP	DET	ADJ	NOUN	1395
PRON	NOUN	ADP	DET	ADJ	NOUN	477

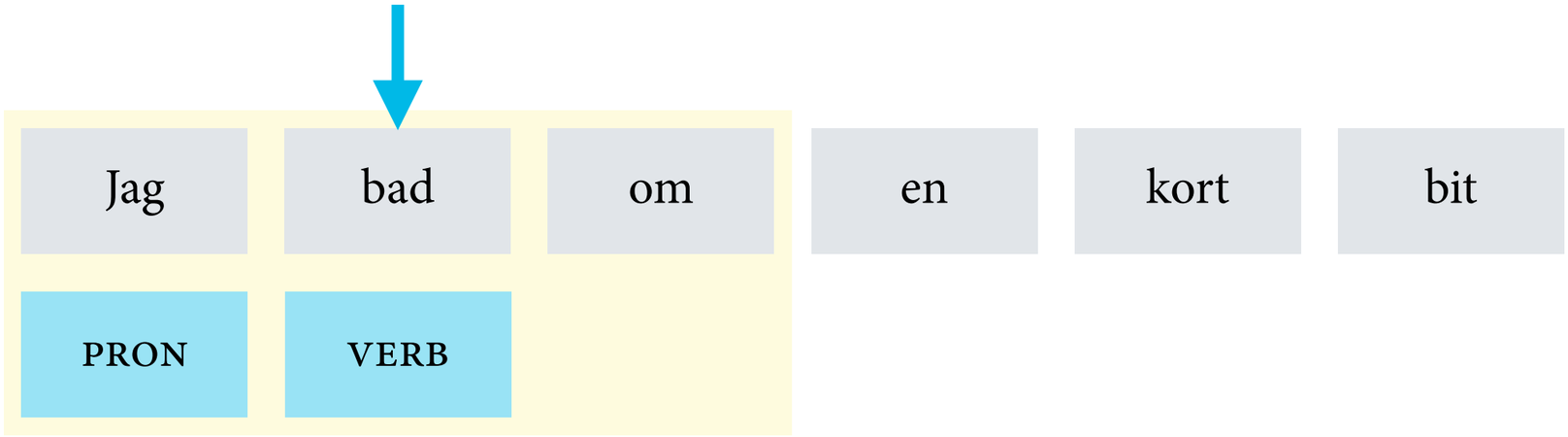
Några kombinationer av taggar
är mer sannolika än andra.

Autoregressiv taggning med fönstermodellen



Modellen predicerar taggen för första ordet.
Klassificeraren ”tittar” på orden innanför fönstret.

Autoregressiv taggning med fönstermodellen



För att predicera nästa tagg kan klassificeraren
använda de redan predicerade taggarna som särdrag.

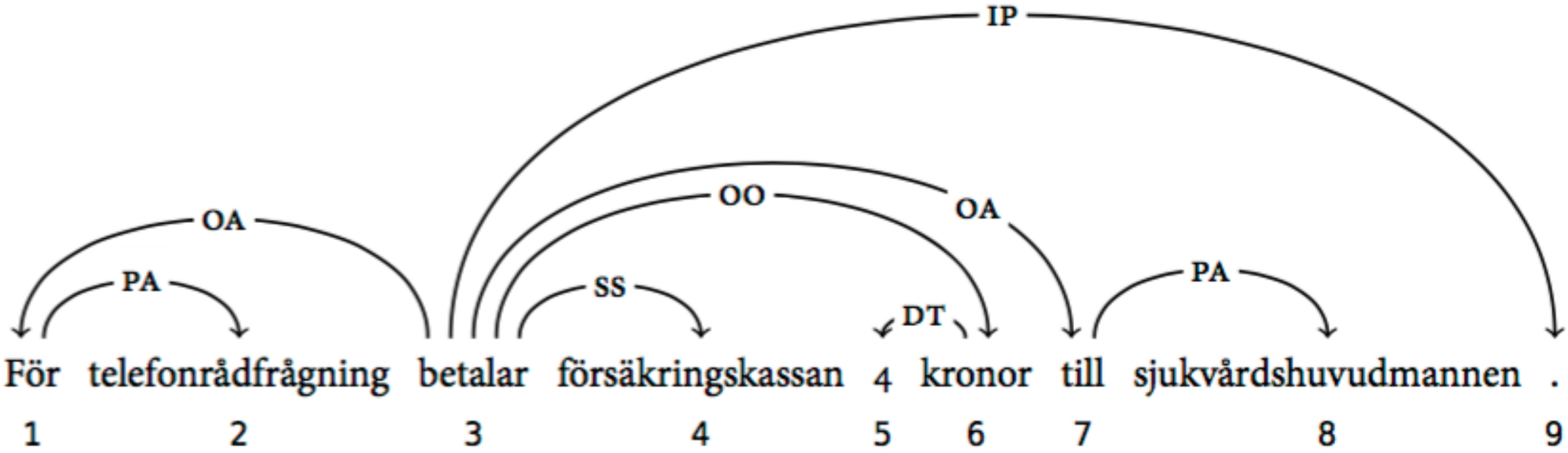
AI för naturligt språk

Dependensträd

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Dependensträd



huvudord ⟶ dependent

Huvudord, dependenter och grammatiska relationer

- Varje båge i ett dependensträd går från ett **huvudord** till ett underordnat ord. Detta ord kallas **dependent**.
- Bågarna är etiketterade med grammatiska relationer såsom subjekt eller objekt.
- Ett dependensträd formaliserar således den information som uttrycks vid traditionell satslösning.

funktionell analys

Hur väljer man huvudord och dependent?

- Huvudordet kan ofta ersätta frasen.

  [Den galne mannen med den vilda blicken] bet den stackars hunden.

- Huvudordet är obligatoriskt; dependenten kan vara valfri.

  Genom skattereformen införs [individuell beskattning].

- Dependentens form beror på huvudordets form.

  Sanne älskar [röda bilar].

AI för naturligt språk

Dependensparsning

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Dependensparsning

- **Dependensparsning** är uppgiften att automatiskt ta fram ett dependensträd för en given mening.
- För att träna en dependensparser krävs data i form av meningar som annoterats med dependensträd.

  trädbanker

- Precis som taggning kan dependensparsning angripas med lokal sökning eller global sökning.

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

Afrikaans	1	49K	IE, Germanic
Akkadian	1	1K	Afro-Asiatic, Semitic
Albanian	1	<1K	IE, Albanian
Amharic	1	10K	Afro-Asiatic, Semitic
Ancient Greek	2	416K	IE, Greek
Arabic	3	1,042K	Afro-Asiatic, Semitic
Armenian	1	52K	IE, Armenian
Assyrian	1	<1K	Afro-Asiatic, Semitic
Bambara	1	13K	Mande
Basque	1	121K	Basque
Belarusian	1	13K	IE, Slavic
Bhojpuri	2	4K	IE, Indic
Breton	1	10K	IE, Celtic
Bulgarian	1	156K	IE, Slavic
Buryat	1	10K	Mongolic
Cantonese	1	13K	Sino-Tibetan
Catalan	1	531K	IE, Romance
Chinese	5	285K	Sino-Tibetan
Classical Chinese	1	74K	Sino-Tibetan
Coptic	1	40K	Afro-Asiatic, Egyptian
Croatian	1	199K	IE, Slavic
Czech	5	2,222K	IE, Slavic
Danish	2	100K	IE, Germanic
Dutch	2	306K	IE, Germanic
English	9	620K	IE, Germanic
Erzya	1	15K	Uralic, Mordvin
Estonian	2	465K	Uralic, Finnic
Faroese	1	10K	IE, Germanic
Finnish	3	377K	Uralic, Finnic
French	8	1,157K	IE, Romance
Galician	2	164K	IE, Romance
German	4	3,753K	IE, Germanic
Gothic	1	55K	IE, Germanic
Greek	1	63K	IE, Greek
Hebrew	1	161K	Afro-Asiatic, Semitic

Google Research Blog

The latest news from Research at Google

Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source

Thursday, May 12, 2016

Posted by Slav Petrov, Senior Staff Research Scientist

At Google, we spend a lot of time thinking about how computer systems can read and understand human language in order to process it in intelligent ways. Today, we are excited to share the fruits of our research with the broader community by releasing SyntaxNet, an open-source neural network framework implemented in TensorFlow that provides a foundation for Natural Language Understanding (NLU) systems. Our release includes all the code needed to train new SyntaxNet models on your own data, as well as *Parsey McParseface*, an English parser that we have trained for you and that you can use to analyze English text.

Utvärdering av dependensparsrar

- **Unlabelled attachment score (UAS)** mäter korrektheten med vilken parsern identifierar rätt huvudord för varje ord.

  liknar korrekthet hos en ordklasstaggare

- **Labelled attachment score (LAS)** räknar poäng endast om både huvudordet *och* den grammatiska relationen är korrekta.

AI för naturligt språk

Transitionsbaserad dependensparsning

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Transitionsbaserad dependensparsning

- En **transitionsbaserad dependensparser** kan ses som en maskin som tar emot instruktioner för att bygga ett dependensträd.
- Parsern börjar i en **initial konfiguration** och avslutar sitt jobb när den har kommit till en **terminal konfiguration**.
- Övergångar mellan konfigurationer kallas **transitioner**.
- Instruktionerna för en transitionsbaserad dependensparser kommer från en klassificerare, t.ex. ett neuronnät.

Konfigurationer

En konfiguration består av tre delar:

- En **buffert,** som innehåller de ord i meningen som fortfarande behöver bearbetas. I början innehåller bufferten alla ord.
- En **stack,** som innehåller de ord i meningen som bearbetas för tillfället. I början är stacken tom.
- Ett **delvist färdigt dependensträd**. I början innehåller det alla ord i meningen men inga dependensbågar.

Transitioner

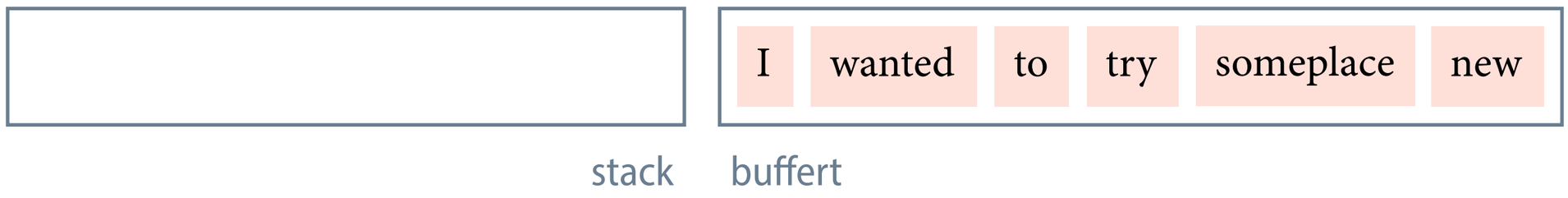
- **Shift (SH)** tar bort det första ordet från bufferten och lägger det på toppen av stacken.

- **Left-arc (LA)** skapar en ny dependensbåge från det översta ordet på stacken till det nästöversta ordet, och tar sedan bort det nästöversta ordet från stacken.

- **Right-arc (RA)** skapar en ny dependensbåge från det nästöversta ordet på stacken till det översta ordet, och tar sedan bort det översta ordet från stacken.

Transitionsbaserad parsning – exempel

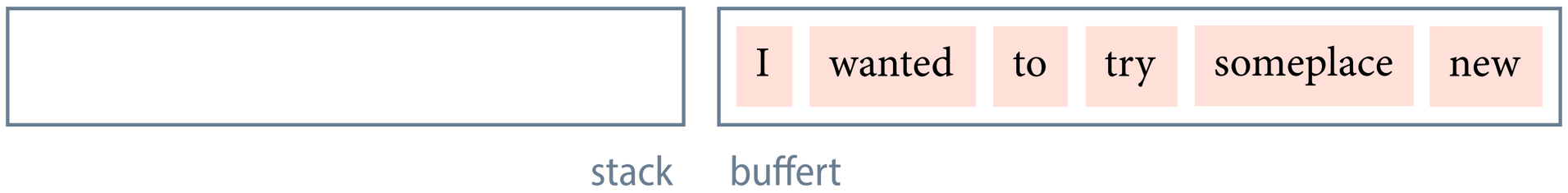
I wanted to try someplace new



(initial konfiguration)

Transitionsbaserad parsning – exempel

I wanted to try someplace new



SH

klassificerare

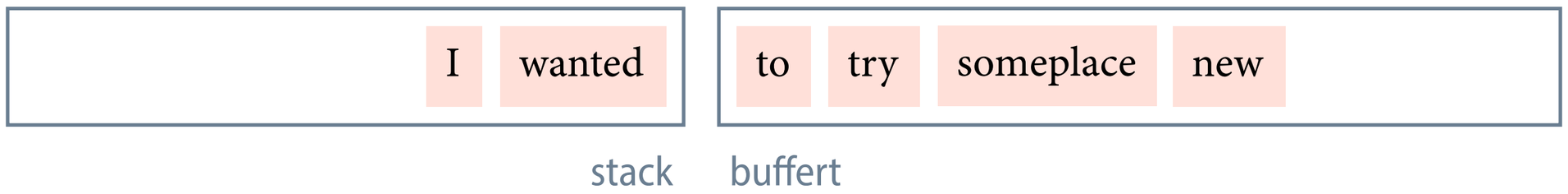
Transitionsbaserad parsning – exempel

I wanted to try someplace new

SH

klassificerare

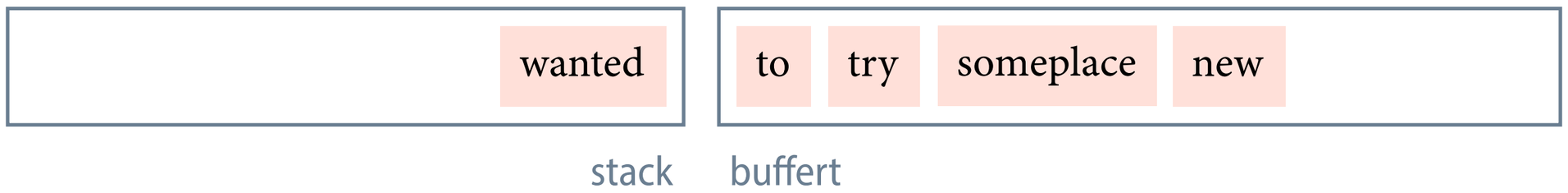
Transitionsbaserad parsning – exempel



LA

klassificerare

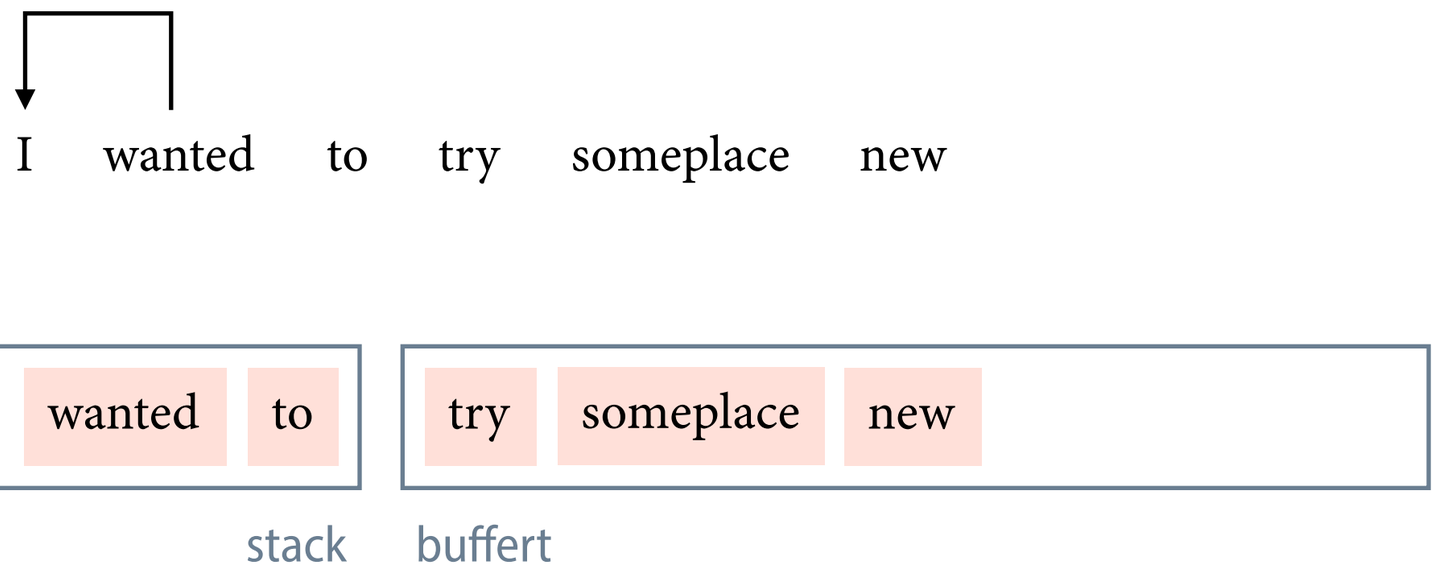
Transitionsbaserad parsning – exempel



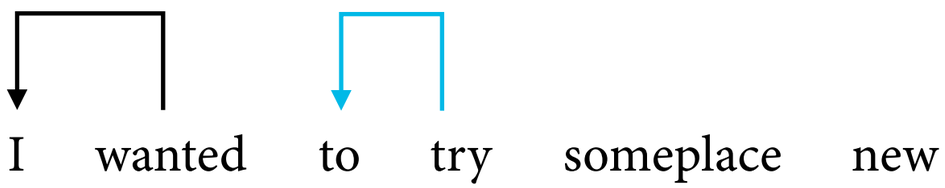
SH

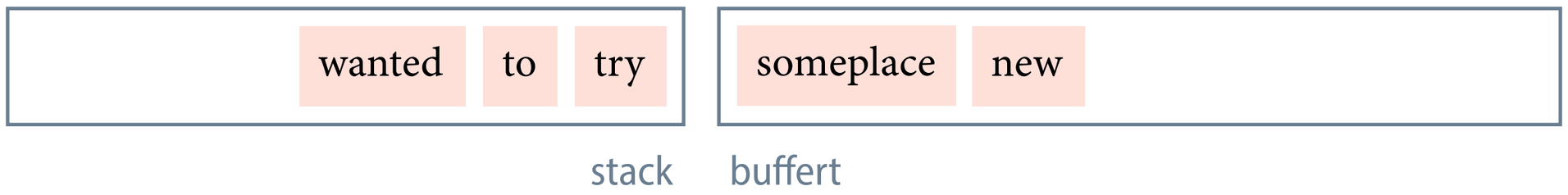
klassificerare

Transitionsbaserad parsning – exempel



Transitionsbaserad parsning – exempel

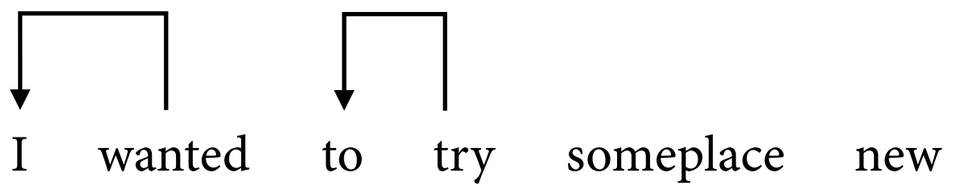


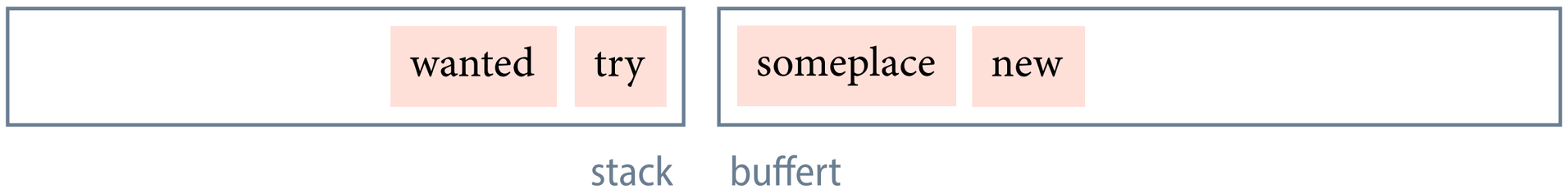


LA

klassificerare

Transitionsbaserad parsning – exempel

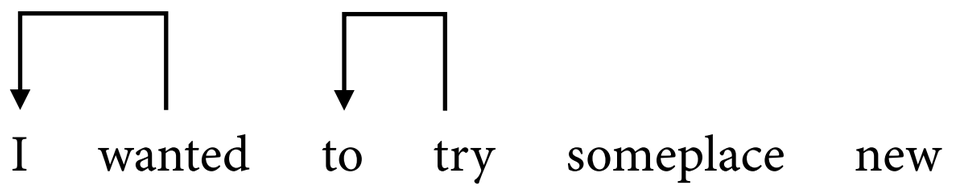


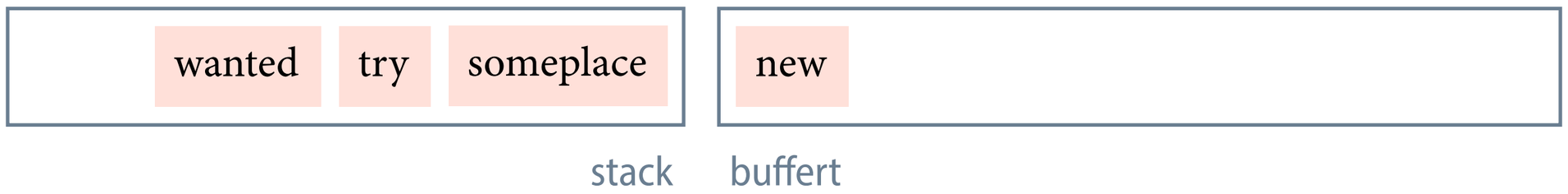


SH

klassificerare

Transitionsbaserad parsning – exempel

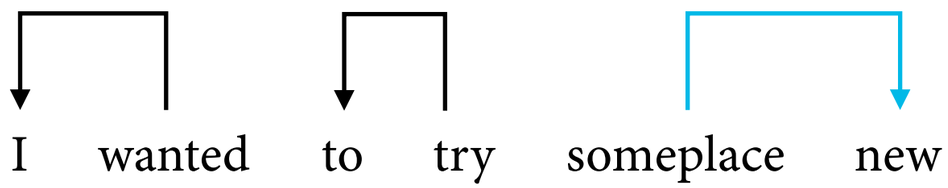


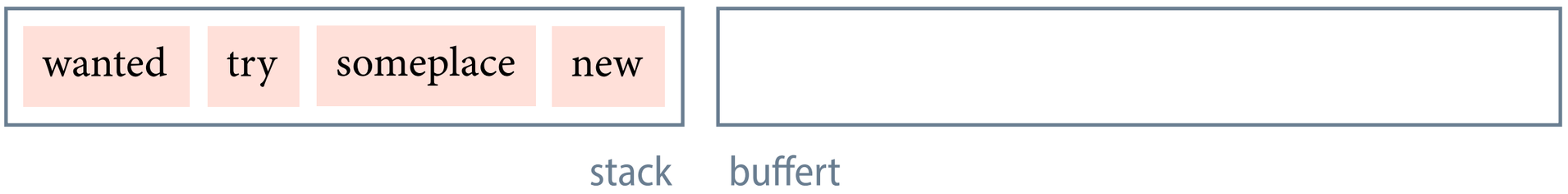


SH

klassificerare

Transitionsbaserad parsning – exempel

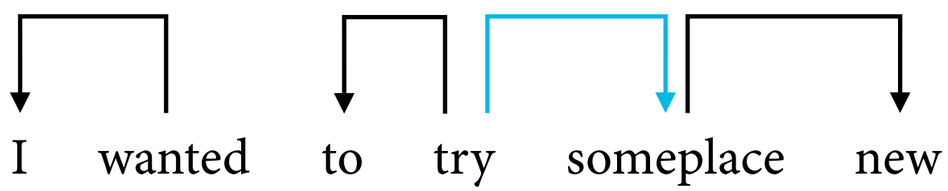




RA

klassificerare

Transitionsbaserad parsning – exempel

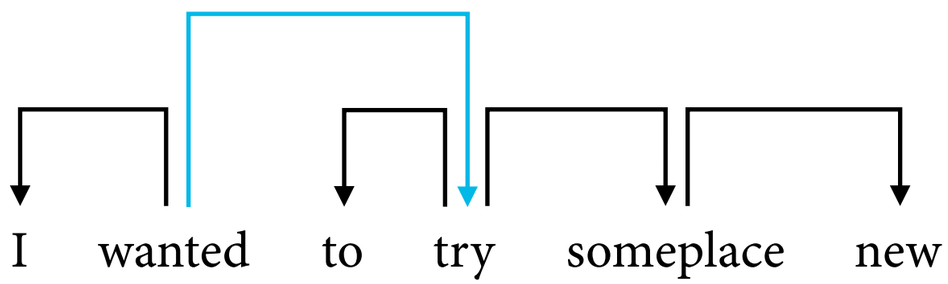


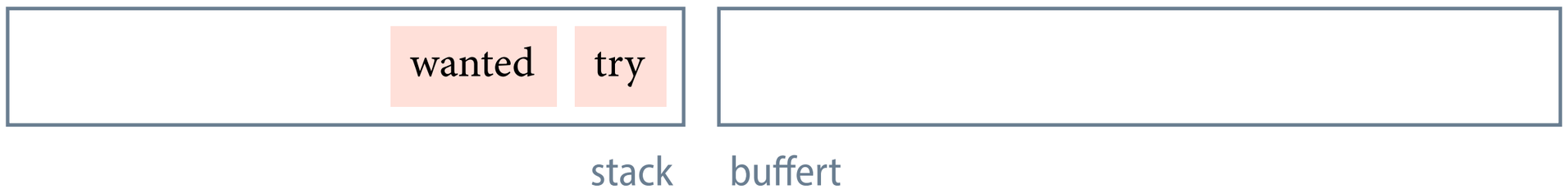
stack buffert

RA

klassificerare

Transitionsbaserad parsning – exempel

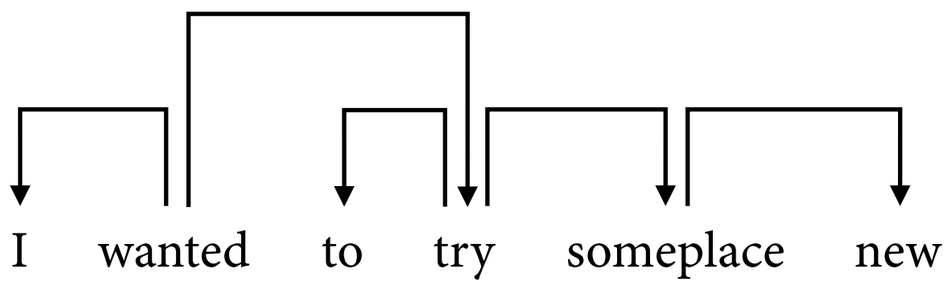


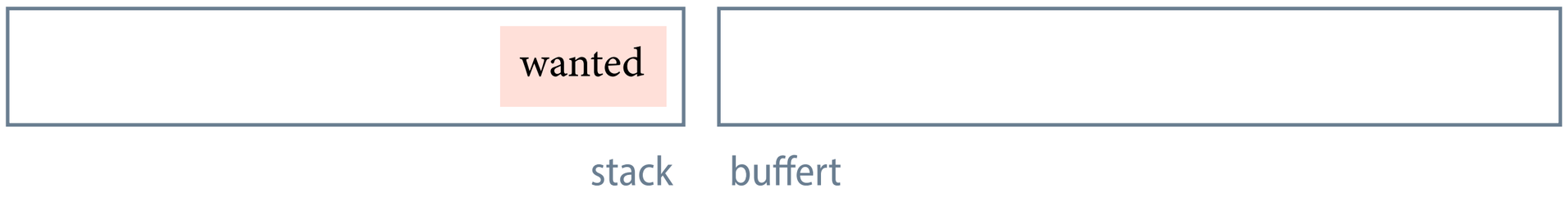


RA

klassificerare

Transitionsbaserad parsning – exempel



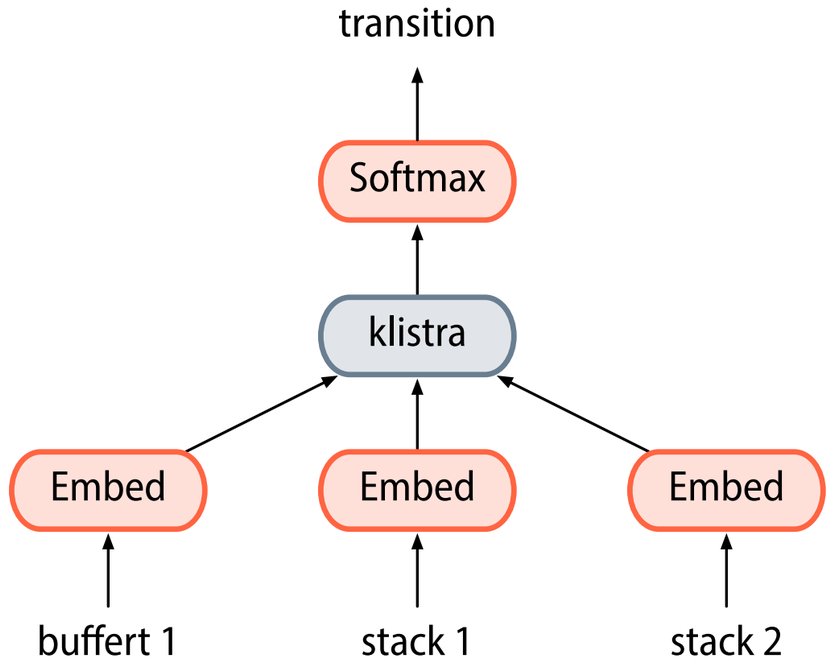


(terminal konfiguration)

Särdrag för klassificeraren

Särdrag för klassificeraren kan definieras över

- orden i bufferten
- orden på stacken
- det delvist färdiga trädet



Chen och Manning (2014)