Applied Natural Language Processing

Info 256
Lecture 24: Information Extraction 1 (April 23, 2019)

David Bamman, UC Berkeley

Announcements

- No office hours this Wed

- Masha will be leading class Thursday

investigating(SEC, Tesla)

SEC Probing Tesla CEO Musk's Tweets: Reports

By Deborah DSouza | Updated August 9, 2018 — 4:47 AM EDT

On Tuesday, Tesla Inc. (TSLA) CEO Elon Musk made the dramatic announcement that he was considering taking Tesla private for $420 a share on Twitter. In an email sent to Tesla employees posted on the company's official blog, Musk explained that he is mulling taking the firm private to protect it from short sellers and wild swings in stock prices. However, the email didn't provide any details regarding financing. (See also: What if Tesla Goes Private?)
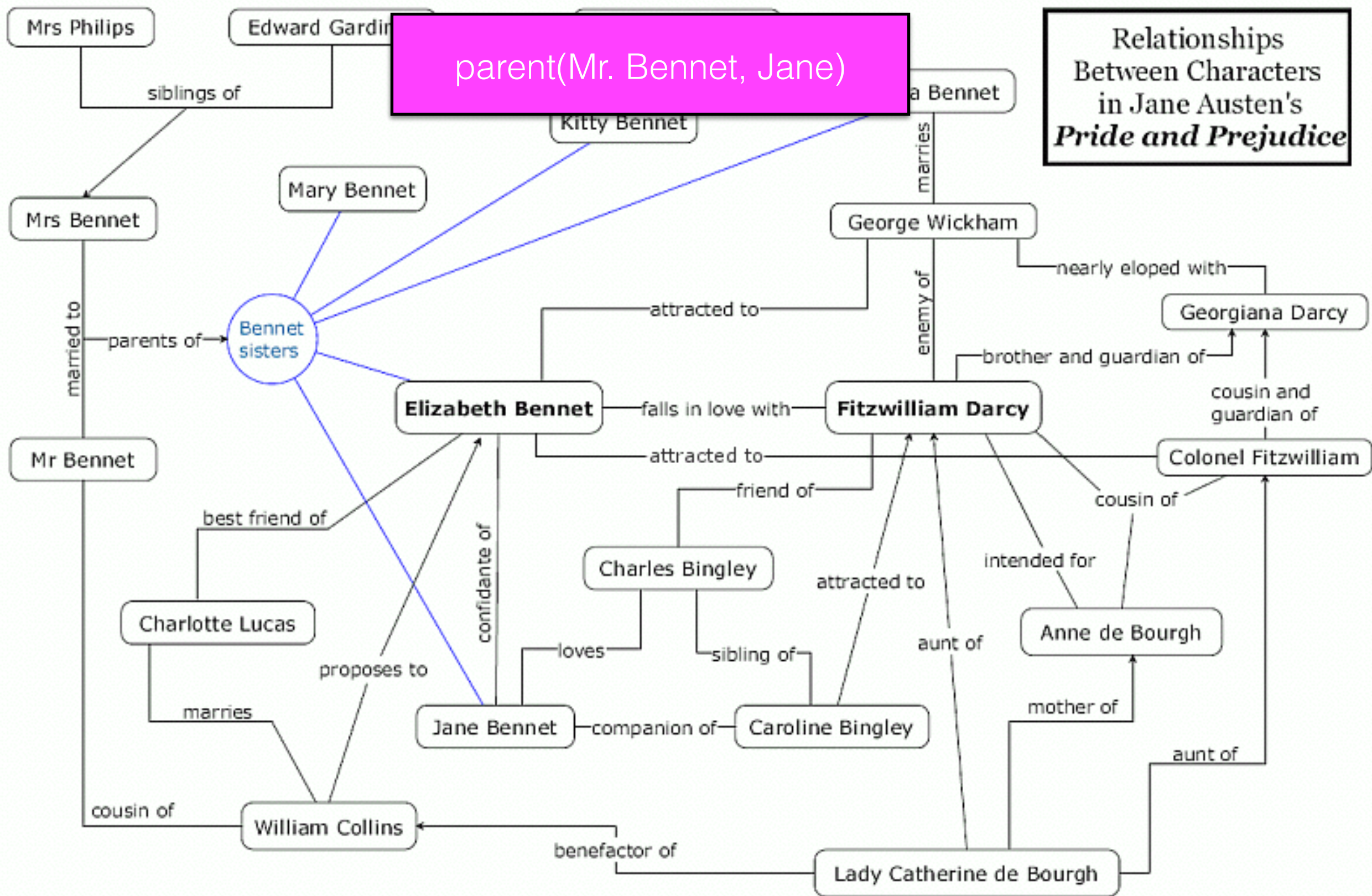
fire(Trump, Sessions)

(Photo: Evan Vucci, AP)

President Donald Trump fired Attorney General Jeff Sessions Wednesday after a year of intense scrutiny from the White House.

Sessions' exit, a resignation forced by Trump, was expected for weeks amid Trump's attacks over his attorney general's decision to recuse himself from the Justice Department's inquiry into Russian meddling in the 2016 election.

Trump also named the attorney general's chief of staff Matthew Whitaker to serve as interim attorney general.

Relationships Between Characters in Jane Austen's *Pride and Prejudice*

parent(Mr. Bennet, Jane)

Mrs Philips

Edward Gardin...

siblings of

Kitty Bennet

...a Bennet

Mary Bennet

Mrs Bennet

married to — parents of →

Bennet sisters

Mr Bennet

George Wickham

marries

enemy of

nearly eloped with

Georgiana Darcy

attracted to

brother and guardian of

cousin and guardian of

Elizabeth Bennet — falls in love with — Fitzwilliam Darcy

attracted to

Colonel Fitzwilliam

friend of

cousin of

best friend of

confidante of

intended for

Charlotte Lucas

Charles Bingley

attracted to

Anne de Bourgh

aunt of

proposes to

marries

loves

sibling of

mother of

cousin of

Jane Bennet — companion of — Caroline Bingley

William Collins

benefactor of

aunt of

Lady Catherine de Bourgh

https://en.wikipedia.org/wiki/Pride_and_Prejudice

Information extraction

- Named entity recognition

- Relation extraction

- Entity linking

Named entity recognition

[tim cook]$_{PER}$ is the ceo of [apple]$_{ORG}$

- Identifying spans of text that correspond to typed entities

Named entity recognition

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Figure 17.1 A list of generic named entity types with the kinds of entities they refer to.

ACE NER categories (+weapon)

# Named entity recognition

- GENIA corpus of MEDLINE abstracts (biomedical)

We have shown that [interleukin-1]PROTEIN ([IL-1]PROTEIN) and [IL-2]PROTEIN control [IL-2 receptor alpha (IL-2R alpha) gene]DNA transcription in [CD4-CD8- murine T lymphocyte precursors]CELL LINE

| protein |
| cell line |
| cell type |
| DNA |
| RNA |

http://www.aclweb.org/anthology/W04-1213

BIO notation

| B-PERS | I-PERS | O | O | O | O | B-ORG |
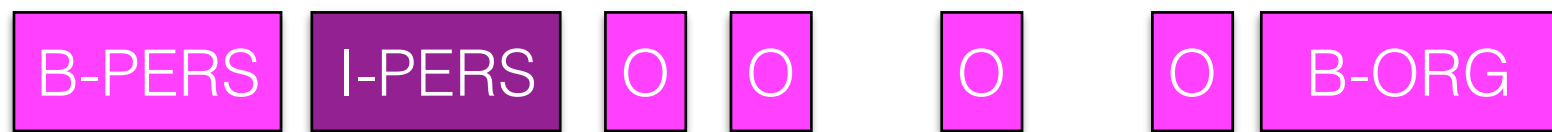
tim cook is the ceo of apple

- **B**eginning of entity
- **I**nside entity
- **O**utside entity

[tim cook]PER is the ceo of [apple]ORG

Relation extraction

The Big Sleep is a 1946 film noir directed by Howard Hawks,[2][3] the first film version of Raymond Chandler's 1939 novel of the same name. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results."[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay.

subject	*predicate*	*object*
The Big Sleep	directed_by	Howard Hawks
The Big Sleep	stars	Humphrey Bogart
The Big Sleep	stars	Lauren Bacall
The Big Sleep	screenplay_by	William Faulkner
The Big Sleep	screenplay_by	Leigh Brackett
The Big Sleep	screenplay_by	Jules Furthman

Relation extraction

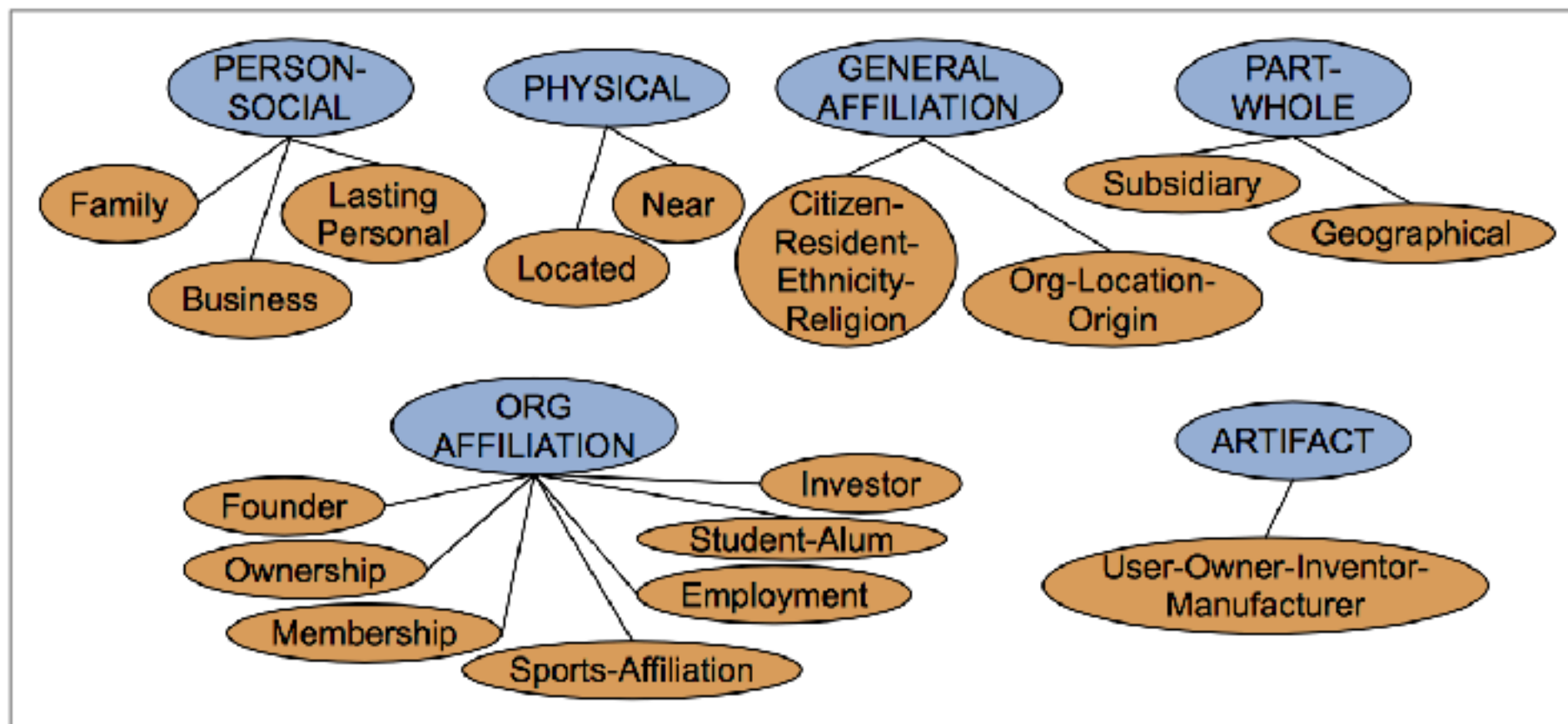**Figure 17.9** The 17 relations used in the ACE relation extraction task.

ACE relations, SLP3

Relation extraction

Entity	Relation	Entity
Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

Unified Medical Language System (UMLS), SLP3

Wikipedia Infoboxes

The Big Sleep

Theatrical release lobby card

The Big Sleep is a 1946 film noir directed by Howard Hawks,[2][3] the first film version of Raymond Chandler's 1939 novel of the same name. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results."[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay.

A remake starring Robert Mitchum as Philip Marlowe was released in 1978. This was the second film in three years featuring Mitchum as Marlowe. The remake was arguably more faithful to the novel, possibly due to fewer restrictions in 1978 on what could be portrayed on screen, however, it was far less successful than the original 1946 version. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.

Directed by	Howard Hawks
Produced by	Howard Hawks
Screenplay by	William Faulkner Leigh Brackett Jules Furthman
Based on	*The Big Sleep* by Raymond Chandler
Starring	Humphrey Bogart Lauren Bacall
Music by	Max Steiner
Cinematography	Sidney Hickox
Edited by	Christian Nyby
Distributed by	Warner Bros.
Release date	August 23, 1946 (United States)
Running time	114 minutes (released cut) 116 minutes (re-released original cut)

Regular expressions

- Regular expressions are precise ways of extracting high-precisions relations

  - "NP$_1$ is a film directed by NP$_2$" → directed_by(NP$_1$, NP$_2$)

  - "NP$_1$ was the director of NP$_2$" → directed_by(NP$_2$, NP$_1$)

Hearst patterns

| pattern	sentence
NP {, NP}* {,} (and|or) other NP$_H$	temples, treasuries, and other important civic buildings
NP$_H$ such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP$_H$ as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP$_H$ {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP$_H$ {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

Hearst 1992; SLP3

Supervised relation extraction

[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$, the first film version of Raymond Chandler's 1939 novel of the same name.

feature(m1, m2)
headwords of m1, m2
bag of words in m1, m2
bag of words between m1, m2
named entity types of m1, m2
syntactic path between m1, m2

Supervised relation extraction

[The Big Sleep]m1 is a 1946 film noir directed by [Howard Hawks]m2, the first film version of Raymond Chandler's 1939 novel of the same name.
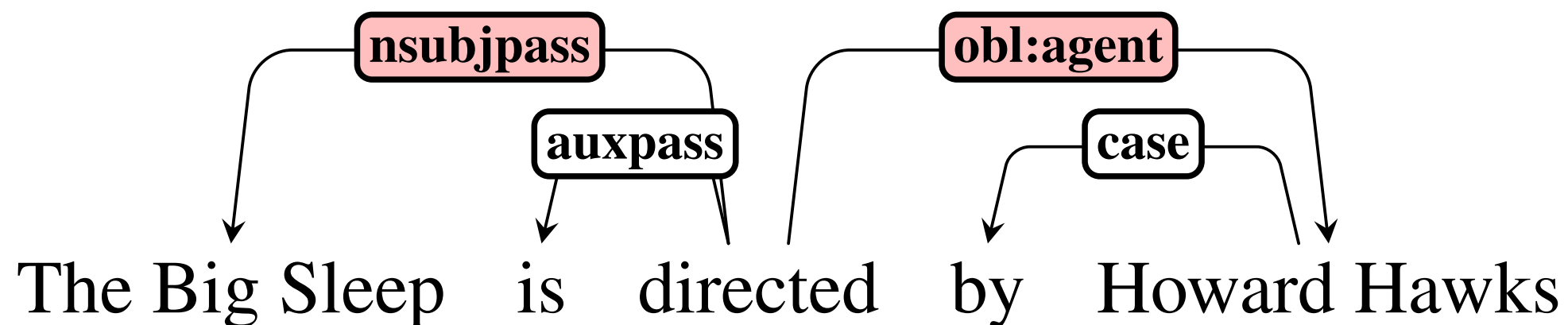
[The Big Sleep]m1 ←*nsubjpass* directed →*obl:agent* [Howard Hawks]m2,

m1←*nsubjpass* ← directed →*obl:agent* → m2

Supervised relation extraction

1.	**George Bush** *traveled to* **France**	George Bush $\xleftarrow{\text{NSUBJ}}$ traveled $\xrightarrow{\text{OBL}}$ France
2.	**Ahab** *traveled to* **Nantucket**	Ahab $\xleftarrow{\text{NSUBJ}}$ traveled $\xrightarrow{\text{OBL}}$ Nantucket
3.	**George Bush** *will travel to* **France**	George Bush $\xleftarrow{\text{NSUBJ}}$ travel $\xrightarrow{\text{OBL}}$ France
4.	**George Bush** *wants to travel to* **France**	George Bush $\xleftarrow{\text{NSUBJ}}$ wants $\xrightarrow{\text{XCOMP}}$ travel $\xrightarrow{\text{OBL}}$ France
5.	**Ahab** *traveled to a city in* **France**	Ahab $\xleftarrow{\text{NSUBJ}}$ traveled $\xrightarrow{\text{OBL}}$ city $\xrightarrow{\text{NMOD}}$ France
6.	*We await* **Ahab** *'s visit to* France	Ahab $\xleftarrow{\text{NMOD:POSS}}$ visit $\xrightarrow{\text{NMOD}}$ France

Supervised relation extraction

```
function FINDRELATIONS(words) returns relations

    relations ← nil
    entities ← FINDENTITIES(words)
    forall entity pairs ⟨e1, e2⟩ in entities do
        if RELATED?(e1, e2)
            relations ← relations + CLASSIFYRELATION(e1, e2)
```

Figure 17.13 Finding and classifying the relations among entities in a text.

directed

max pooling layer

convolutional layer

...

word embedding
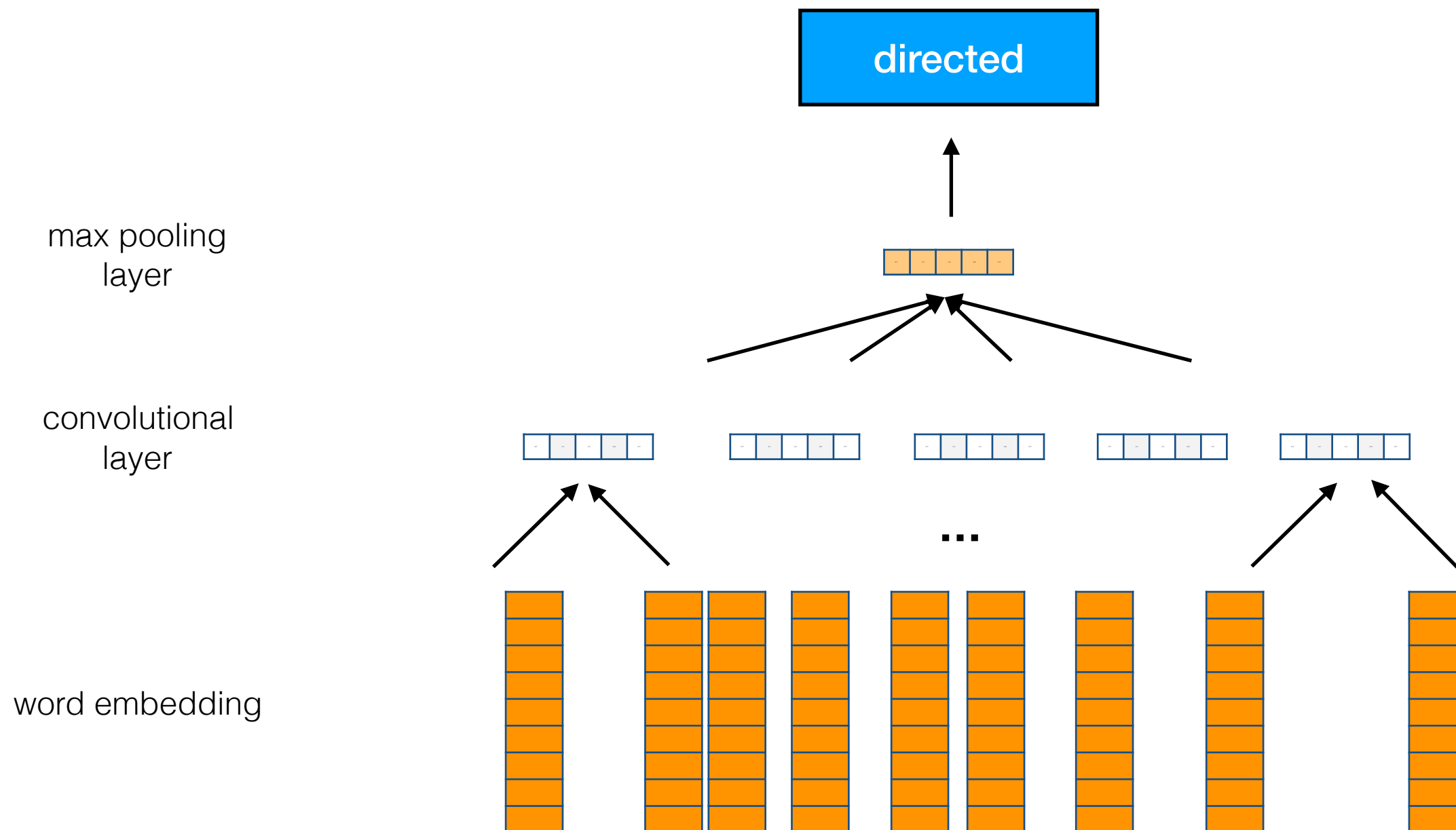
[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$

We don't know which entities we're classifying!

directed(Howard Hawks, The Big Sleep)
genre(The Big Sleep, Film Noir)
year_of_release(The Big Sleep, 1946)

Neural RE

- To solve this, we'll add positional embeddings to our representation of each word — the distance from each word w in the sentence to m1 and m2
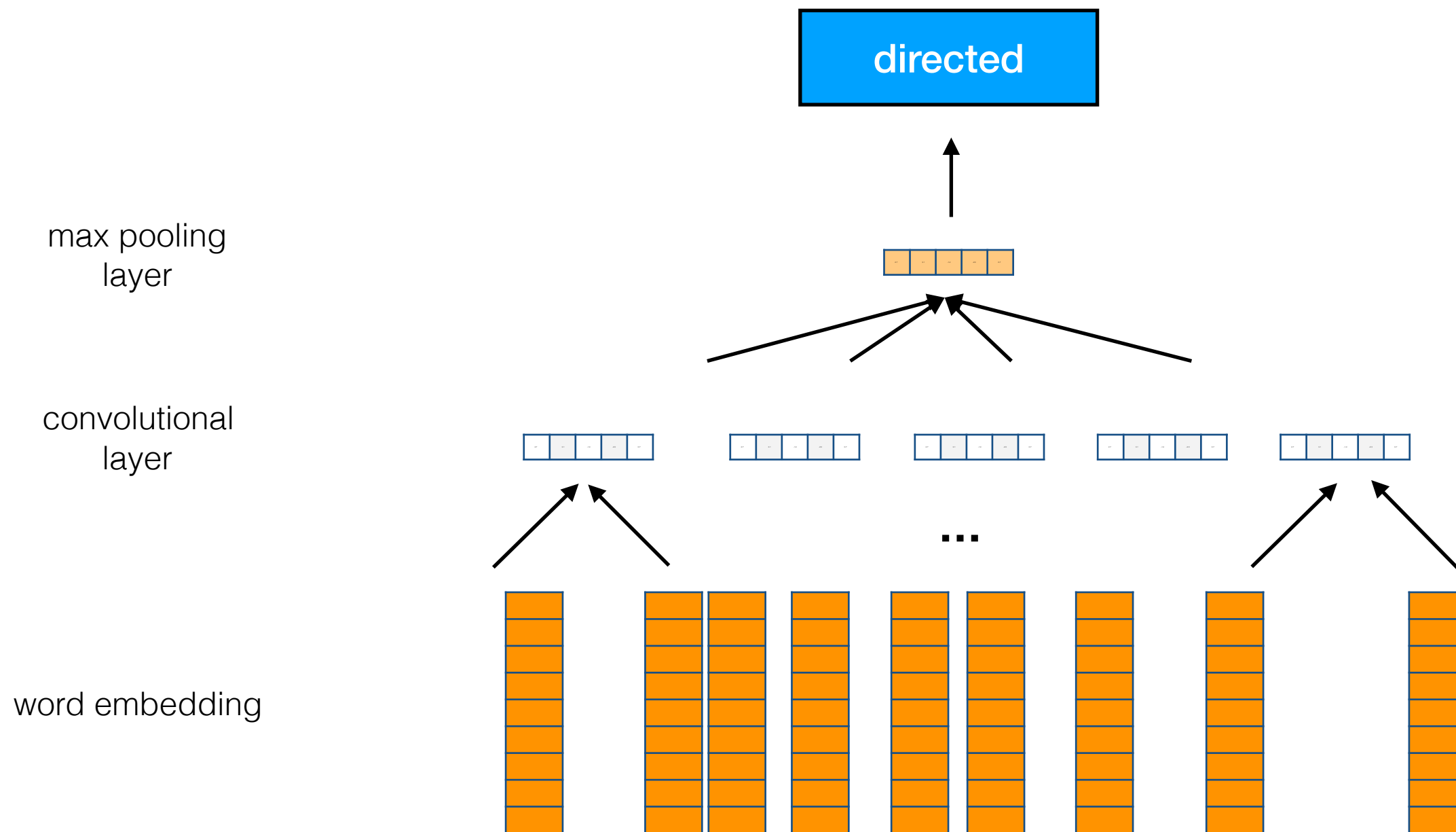
dist from m1	0		1	3	4	5	6	7	8	9
dist from m2	-8		-7	-6	-5	-4	-3	-2	-1	0
	[The Big Sleep]	is	a	1946	film	noir	directed	by	[Howard Hawks]	

- 0 here uniquely identifies the head and tail of the relation; other position indicate how close the word is (maybe closer words matter more)
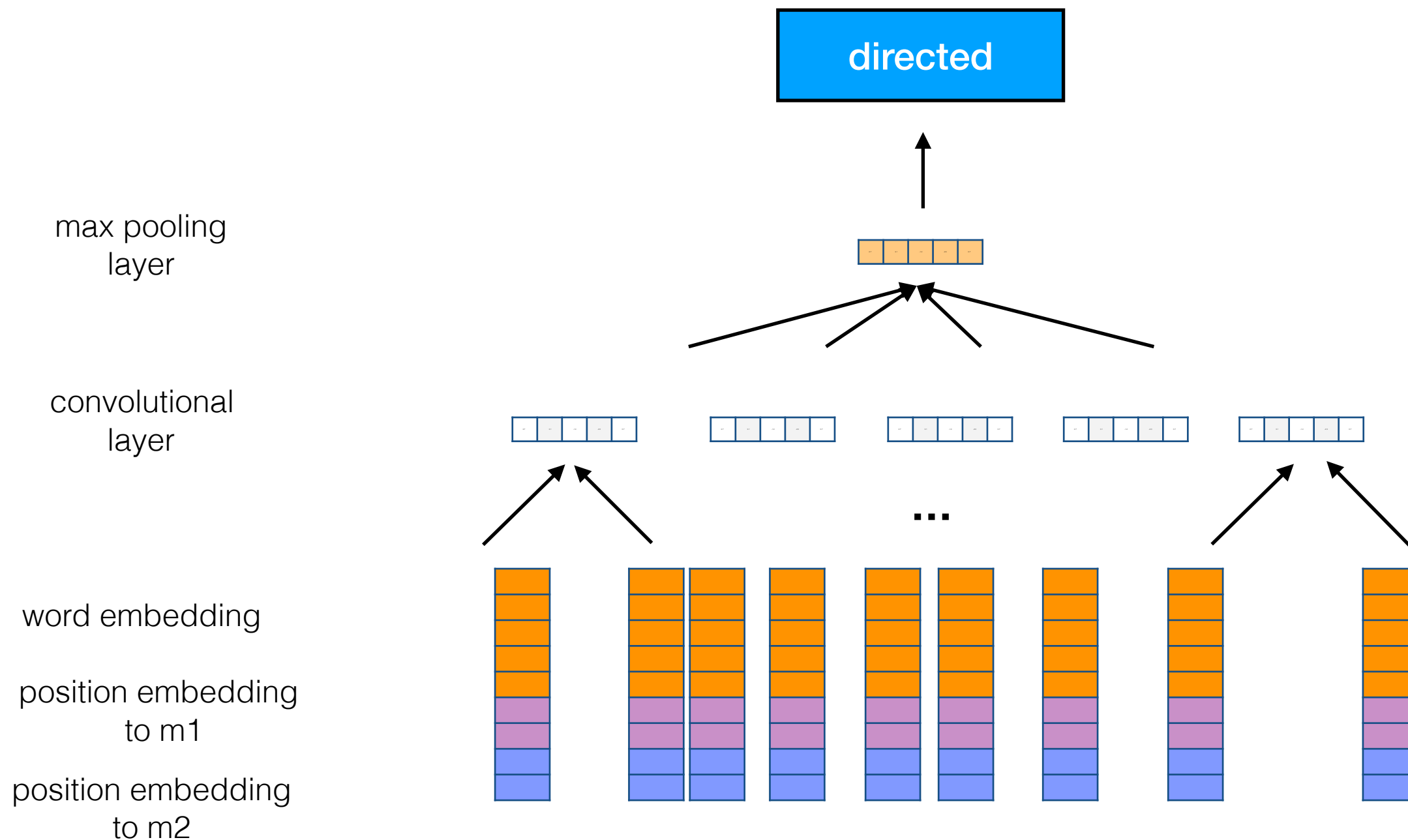
Neural RE

Each position then has an embedding

-4	2	-0.5	1.1	0.3	0.4	-0.5
-3	-1.4	0.4	-0.2	-0.9	0.5	0.9
-2	-1.1	-0.2	-0.5	0.2	-0.8	0
-1	0.7	-0.3	1.5	-0.3	-0.4	0.1
0	-0.8	1.2	1	-0.7	-1	-0.4
1	0	0.3	-0.3	-0.9	0.2	1.4
2	0.8	0.8	-0.4	-1.4	1.2	-0.9
3	1.6	0.4	-1.1	0.7	0.1	1.6
4	1.2	-0.2	1.3	-0.4	0.3	-1.0

directed

max pooling
layer

convolutional
layer

word embedding

...

[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$

max pooling
layer

convolutional
layer

word embedding

position embedding
to m1

position embedding
to m2

directed

...

[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$

Distant supervision

- It's uncommon to have labeled data in the form of <sentence, relation> pairs

sentence	relations
[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$, the first film version of Raymond Chandler's 1939 novel of the same name.	directed_by(The Big Sleep, Howard Hawks)

Distant supervision

- More common to have knowledge base data about entities and their relations that's separate from text.

- We know the text likely expresses the relations somewhere, but not exactly where.

Wikipedia Infoboxes

The Big Sleep

Theatrical release lobby card

The Big Sleep is a 1946 film noir directed by Howard Hawks,[2][3] the first film version of Raymond Chandler's 1939 novel of the same name. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results."[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay.

A remake starring Robert Mitchum as Philip Marlowe was released in 1978. This was the second film in three years featuring Mitchum as Marlowe. The remake was arguably more faithful to the novel, possibly due to fewer restrictions in 1978 on what could be portrayed on screen, however, it was far less successful than the original 1946 version. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.

Directed by	Howard Hawks
Produced by	Howard Hawks
Screenplay by	William Faulkner Leigh Brackett Jules Furthman
Based on	*The Big Sleep* by Raymond Chandler
Starring	Humphrey Bogart Lauren Bacall
Music by	Max Steiner
Cinematography	Sidney Hickox
Edited by	Christian Nyby
Distributed by	Warner Bros.
Release date	August 23, 1946 (United States)
Running time	114 minutes (released cut) 116 minutes (re-released original cut)

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Table 2: The 23 largest Freebase relations we use, with their size and an instance of each relation.

Mintz et al. 2009

Distant supervision

mayor(Maynard Jackson, Atlanta)

Elected mayor of Atlanta in 1973, Maynard Jackson…
Atlanta's airport will be renamed to honor Maynard Jackson, the city's first Black mayor
Born in Dallas, Texas in 1938, Maynard Holbrook Jackson, Jr. moved to Atlanta when he was 8.

mayor(Fiorello LaGuardia, New York)

Fiorello LaGuardia was Mayor of New York for three terms...
Fiorello LaGuardia, then serving on the New York City Board of Aldermen...

Eisenstein 2018

Distant supervision

- For feature-based models, we can represent the tuple <m1, m2> by aggregating together the representations from all the sentences they appear in

Distant supervision

[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$, the first film version of Raymond Chandler's 1939 novel of the same name.

[Howard Hawks]$_{m2}$ directed the [The Big Sleep]$_{m1}$

feature(m1, m2)	value (e.g., normalized over all sentences)
"directed" between m1, m2	0.37
"by" between m1, m2	0.42
m1 ← *nsubjpass* ← directed → *obl:agent* → m2	0.13
m2 ← *nsubj* ← directed → *obj* → m2	0.08

Distant supervision

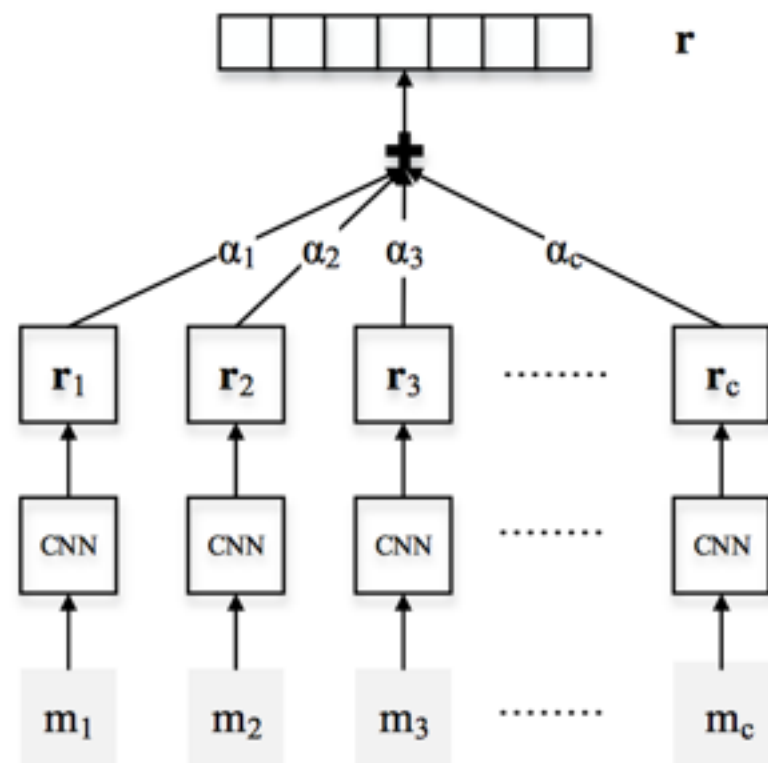- Discovering Hearst patterns from distant supervision using WordNet (Snow et al. 2005)

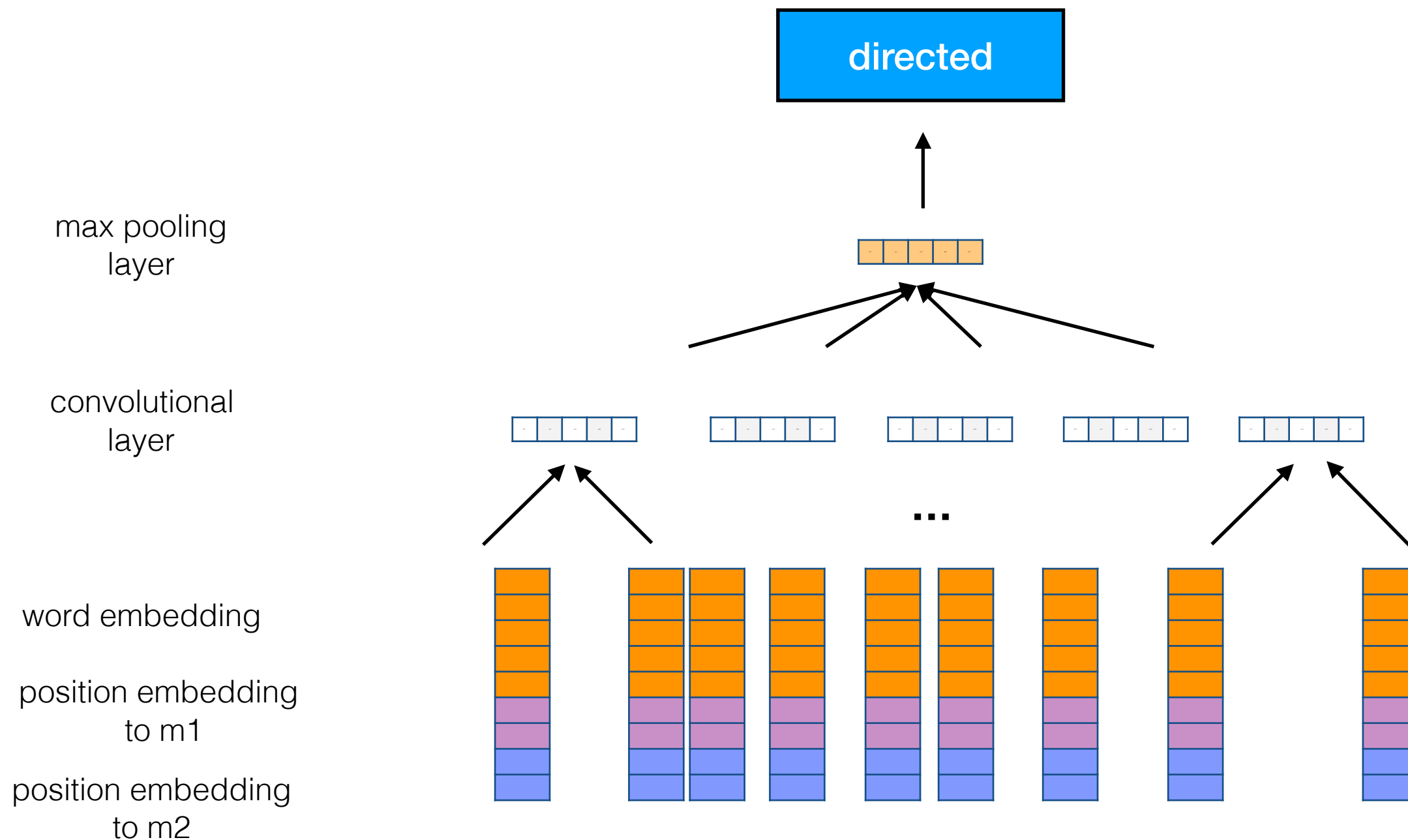| pattern	sentence
NP_H like NP	Many hormones like leptin...
NP_H called NP	a markup language called XHTML
NP is a NP_H	Ruby is a programming language...
NP, a NP_H	IBM, a company with a long...

SLP3

Multiple Instance Learning

- Labels are assigned to a set of sentences, each containing the pair of entities m1 and m2; not all of those sentences express the relation between m1 and m2.

Attention

- Let's incorporate structure (and parameters) into a network that captures which sentences in the input we should be attending to (and which we can ignore).
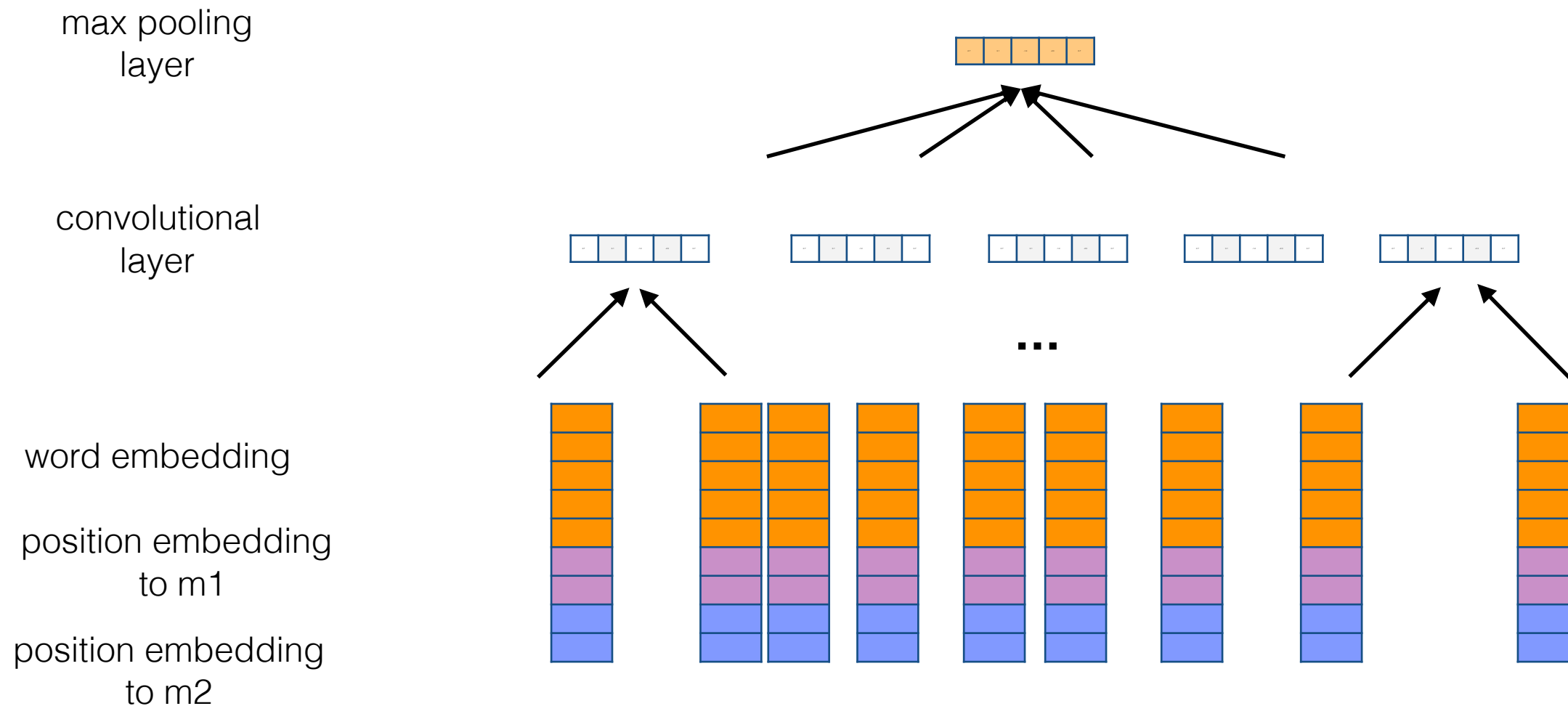

Lin et al (2016), "Neural Relation Extraction with Selective Attention over Instances" (ACL)

directed

max pooling layer

convolutional layer

word embedding

position embedding to m1

position embedding to m2

...

[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$

Lin et al (2016), "Neural Relation Extraction with Selective Attention over Instances" (ACL)

Now we just have an encoding of a sentence

max pooling layer

convolutional layer

word embedding

position embedding to m1

position embedding to m2

[The Big Sleep]m1 is a 1946 film noir directed by [Howard Hawks]m2

Lin et al (2016), "Neural Relation Extraction with Selective Attention over Instances" (ACL)

directed

$$x_1a_1 + x_2a_2 + x_3a_3$$

weighted sum

| 2.7 | 3.1 | -1.4 | -2.3 | 0.7 |

sentence encoding

| 2.7 | 3.1 | -1.4 | -2.3 | 0.7 |

| 2.7 | 3.1 | -1.4 | -2.3 | 0.7 |

| 2.7 | 3.1 | -1.4 | -2.3 | 0.7 |

[The Big Sleep]m1 is a 1946 film noir directed by [Howard Hawks]m2

[Howard Hawks]m2 directed [The Big Sleep]m1

After [The Big Sleep]m1 [Howard Hawks]m2 married Dee Hartford
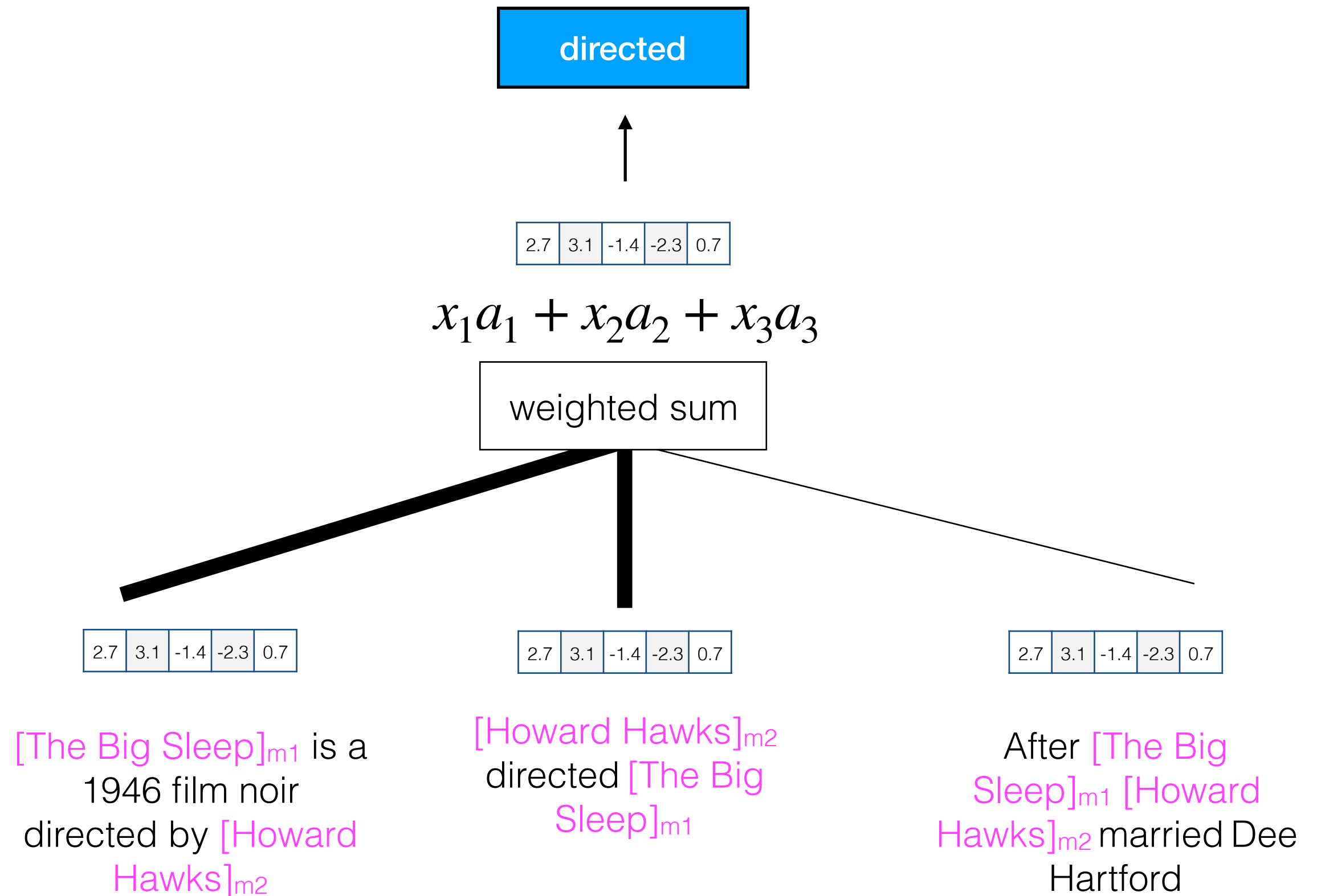
Information Extraction

- Named entity recognition

- Entity linking

- Relation extraction

- Templated filling

- Event detection

- Event coreference

- Extra-propositional information (veridicality, hedging)

Activity

- 16.ie/DependencyPatterns_TODO.ipynb