Natural Language Processing

Lecture 4
Semantic analysis

The study of semantics within the theory "Meaning ⇔ Text"

A. Melchuck: one has to build a formal model, which covers all grammatical phenomena of a particular language. Describing of the text interpretation process (meaning identifying) may be acquired on the base of describing of the text building process.

Lexical function

Lexical function is a function, arguments of which are words or phrases of this language, values are set of words or phrases of this language.

The values of one function with different arguments may overlap; The values of different functions with the same argument may overlap:

$$\exists f_1 \neq f_2 \exists x: f_1(x) = f_2(x);$$

$$\exists f \exists x_1 \neq x_2 : f(x_1) = f(x_2)$$

Lexical function

Example:

The value of the predicate Syn(x, y) is true if x, y are synonyms; false in other cases.

The value of the predicate Anti(x, y) is true if x, y are antonyms; false in other cases.

Gener(x, y): y is generalizing concept regarding to the concept denoted as x (x = strawberry, y = berry).

Situation

Situation is a lexical reflection in this language of some parts of reality.

Semantic actants (meaning components): A, B, C, D.

Deep syntactic actants (dependent): 1, 2, 3, 4.

Loc(x, y), y is a preposition of localization (spatial, temporal or abstract).

 $Loc_{in}(x, y)$, y is a "statistic" localization (in Moscow).

 $Loc_{ad}(x, y)$, y is a preposition of direction (to Moscow).

 $Loc_{ab}(x, y)$, y is a preposition of distance (from Moscow).

 $Func_1(x, y)$, y is a verb, linking a situation as a subject with actants as objects (*Sun* is shining through the window.).

Labor₁₂(x, y): "The president awarded him the order".

Perf(x, y), y is a completed action (x = читать, y = прочитать).

Destr(x, y), y is an "aggressive" action (x = bee, y = bite).

Cap(x, y), y is a "chef" (x = department, y = dean).

Valency

Word valency is ability of words to form links with other words.

Semantic valencies are defined by the lexical analysis of a situation, defined by the particular word.

To rent: $P(x_1, x_2, x_3, x_4, x_5)$, where x_1 - "who", x_2 - "what", x_3 - "from whom", x_4 - "price", x_5 - "period".

Syntactic valencies are defined by the number of syntactic actants, which are presented in the text and depend on a context.

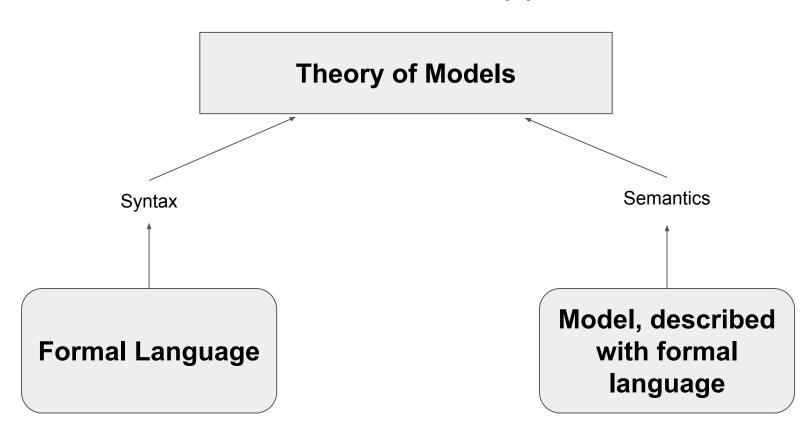
Explanatory Combinatorial Dictionary

The model, which was suggested by I.A.Melchuck, presents a language as a set of dictionary entries with a huge amount of diverse information.

Dictionary entry contains:

- basic word;
- lexical predicates;
- valencies.

Theoretical Model Approach



Signature is a set $\Sigma = \langle \Sigma_R, \Sigma_F, \Sigma_C, \rho \rangle$ consisting of:

- Σ_R symbols for relations (predicates);
- Σ_{E} symbols for operations (functions);
- $\Sigma_{\rm C}$ symbols for constants;
- ρ function, which associates names of predicates and operations with their arity: $\rho: \Sigma_R \cup \Sigma_F \to \mathbb{N}$.

Alphabet of a formal language consists of :

- signature symbols: Σ_R , Σ_F , Σ_C , ρ ;
- variable symbols $(x, y, z, u, w... \in V)$;
- logical connective (∧, ∨, →, ¬);
- ∀ and ∃;
- additional symbols: =, (,), ,.

Term of signature Σ with variables from V is defined by induction:

- 1. Basis of induction. Any variable x ($x \in V$) is term c, and any constant symbol $c \in \Sigma_c$ of fixed signature Σ_c is term c as well;
- 2. Step of induction. Let terms $t_1, ..., t_m$ be already built, f is a symbol of operation $(f \in \Sigma_F)$, and arity of operation is equal to m (p(f) = m), then $f(t_1, ..., t_m)$ is a term as well.

Formula of signature Σ with variables from V is defined by induction:

- 1. Basis of induction. If p, q are terms, then expression p = q is a formula, and for any variable x ($x \in V$) x is a free variable in this formula; there isn't any bound variables. If P is symbol of n-ary predicate of signature Σ (p(P) = n), and $t_1, ..., t_n$ are terms of this signature, then the expression $P(t_1, ..., t_n)$ is a formula, and for any variable x ($x \in V$) x is a free variable in this formula; there isn't any bound variables. These formulae are called **atomic**.
- 2. Step of induction. Let formulae φ and ψ be already built, then the expressions $\varphi \land \psi$, $\varphi \lor \psi$, $\varphi \to \psi$, $\neg \varphi$ are formulae. Subformulae of these formulae are these formulae and all subformulae of formulae φ and ψ . Subformulae of formula $\neg \varphi$ is this formula and all subformulae of formula φ .

If φ is a formula, and x is a variable, then $(\forall x)\varphi$ and $(\exists x)\varphi$ are defined.

Sentence (or close formula) is a formula without free variables.

First-order logic is any set of sentences.

Axioms of the first-order logic:

2.
$$((A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C)));$$

10. $(A \rightarrow C) \rightarrow ((B \rightarrow C) \rightarrow ((A \lor B) \rightarrow C))$:

3.
$$A \wedge B \rightarrow A$$
;

3.
$$A \wedge B \rightarrow A$$
;
4. $A \wedge B \rightarrow B$:

5.
$$A \rightarrow (B \rightarrow (A \land B))$$
;

1. $A \rightarrow (B \rightarrow A)$:

6. $A \rightarrow (A \lor B)$:

7. $B \rightarrow (A \lor B)$;

9. A V ¬A:

8. $\neg A \rightarrow (A \rightarrow B)$;

12. $\forall xA \rightarrow A[t/x]$;

13. A[t / x] $\rightarrow \exists xA$.



11. $(A \rightarrow B) \rightarrow ((A \rightarrow \neg B) \rightarrow \neg A)$





















Rules of the first-order logic:

$$\frac{A, A \to B}{B} \qquad \qquad \frac{A}{\forall xA}$$

Characteristics of the first-order logic:

- Consistency: no formula can be deduced simultaneously with its negation;
- Completeness: for any sentence of the logic it is deduced or its negation is deduced (Gödel's completeness theorem).

Example:

"If I work hard, I'll get a reward. If I get a reward, I'll by a bike. Consequently, if I work hard, I'll buy a bike."

A = "I'll work hard";

B = "I'll get a reward";

C = "I'll buy a bike".

 $((A \rightarrow B) \land (B \rightarrow C)) \rightarrow (A \rightarrow C).$

Set-theoretic approach (S. Marcus, 1970)

Vocabulary Γ is a finite set of words.

Consider free semigroup T on Γ , i.e. a set of all finite sequences of words with defined associative and noncommutative operation of concatenation. Semigroup T is called **complete** (or **universal language**) on Γ .

Null sequence (denoted as θ) is such sequence that $\theta x = x\theta = x$ for each sequence x.

A pair $\{\Gamma, \Phi\}$ is called **language** on Γ , where $\Phi \subseteq T$.

Two words a and b belong to the same **distributive class**, if for each pair of sequences x, y the condition $xay \in \Phi \leftrightarrow xby \in \Phi$ is fulfilled, i.e. a sequence $xay \in \Phi$ iff $xby \in \Phi$.

Context is an ordered pair of sequences on Γ (denoted as $\langle x, y \rangle$, where $x, y \in T$).

Word a is accepted in context $\langle x, y \rangle$ if sequence $xay \in \Phi$.

Denote a set of all contexts as J(a), in which word a is accepted. Then two words a and b belong to the same distributive class iff J(a) = J(b), i.e. a and b are accepted in the same contexts.

Distributive division, containing word a, is denoted as S(a).

Language is a triple $\{\Gamma, P, \Phi\}$, where P is dictionary Γ division, different from the division on distributive classes. If P is a division of Γ , then each subset of P is called a cell from P (or P-cell).

If a division P is defined, then $\Gamma = \bigcup P_i$, where P_i is a cell from P, n is a number of cells.

Denote a cell from P as P(a), which contains word a. For two different words a and b P(a) = P(b) or $P(a) \cap P(b) = \emptyset$.

P(a) is interpreted as a set of inflected forms of word a. Singular division of Γ : each cell consists of one word.

Consider two divisions P and Q of vocabulary Γ . Division P is smaller than division Q, if $P(a) \subseteq Q(a)$ for each $a \in \Gamma$.

If $x_1x_2...x_n$ is sequence on Γ , then the sequence $P(x_1)P(x_2)...P(x_n)$ is called **P-structure** of sequence $x_1x_2...x_n$.

P is **regularly smaller** than *Q* if *P* is smaller than *Q*, and for each triple of words *x*, *y*, *z* the following conditions are fulfilled: $P(x) \subseteq Q(z)$ and $P(y) \subseteq Q(z)$ brings *P*-equivalence $P(x) \leftrightarrow P(y)$.

For each division P on Γ consider division P, which cells are defined as follows: $P'(x) = UP(y) [P(x) \leftrightarrow P(y)]$ (for each $x \in \Gamma$). Division P is called derivative from division P.

P is regularly smaller than *P*'.

Language is a triple $\{\Gamma, P, \Phi\}$, where Γ is a finite vocabulary, P is a division of vocabulary Γ , which is different from distributive one, Φ is a subset of free semigroup on Γ .

The **chain between words** a and b is a finite sequence $x_1, x_2, ..., x_i, x_{i+1}, x_n$ such that $x_1 = a$, $x_n = b$ and $x_i \in S(x_{i+1}) \cup P(x_{i+1})$ for $1 < i \le n-1$.

R(a) is a set of words **b**, for each of them the chain exists, connecting a and b.

- $a \in R(a)$;
- if $b \in R(a)$ then $a \in R(b)$;
- if $c \in R(b)$ and $b \in R(a)$, then $c \in R(a)$.

For each $a \in \Gamma$: $S(a) \subseteq R(a)$ and $P(a) \subseteq R(a)$.

Denote as K(a) a set of such words b that at least one of the following conditions are fulfilled:

- 1. $S(a) \cap P(b) \neq \emptyset$;
- 2. $S(b) \cap P(a) \neq \emptyset$.

Set K(a) is called **class** a.

Since $a \in S(a) \cap P(a)$, then $a \in K(a)$ for each $a \in \Gamma$.

Language $\{\Gamma, P, \Phi\}$ is called **adequate** if for each $x \in \Gamma$: $S(x) \subseteq P'(x)$.

Language $\{\Gamma, P, \Phi\}$ is called **homogeneous** if from $S(x) \cap P(y) \neq \emptyset$ $(x \in \Gamma, y \in \Gamma)$ one can deduce $S(y) \cap P(x) \neq \emptyset$.

Consider language $\{\Gamma, P, \Phi\}$, where

- Γ vocabulary of natural language L;
- P(x) ($x \in \Gamma$) set of all forms of word x;
- Φ set of all correct sentences in language L.

If word x has two different forms x_1 and x_2 then $P(x_1) \cap P(x_2) = \emptyset$.

For each natural language: if two words a and b belong to the same paradigm (i.e. $b \in P(a)$), then they have the same part of speech. Thus, "part of speech" is join of P-cells.

Given two words a and b. Cells P(a) and P(b) associate with the same **part of** speech iff P(a) and P(b) are P-equivalent, i.e. iff $b \in P'(a)$.

Let $\{\Gamma, P, \Phi\}$ be a language. Two words - $a \in \Gamma$ and $b \in \Gamma$ - belong to the same **grammatical gender** (denoted as $a\gamma b$) if for each $a' \in P(a)$ and each $b' \in P(b)$ at least one of the following conditions is fulfilled: $P(a) \cap S(b') \neq \emptyset$, $P(b) \cap S(a') \neq \emptyset$.

Theorem 1. Non-adequate language exists.

Theorem 2. If $\{\Gamma, P, \Phi\}$ is adequate language then R' = P'.

<u>Theorem 3</u>. Let $\{\Gamma, P, \Phi\}$ be adequate language. If classes P(x) define the division K of vocabulary Γ , then K' = P'.

<u>Theorem 4</u>. If language $\{\Gamma, P, \Phi\}$ is homogeneous, then K(x) = R(x) for each $x \in \Gamma$.

Consider a sentence "Счастливые дети бегают во дворе и веселятся".

Consider a sentence "Happy children run in the garden and have fun."

b = children; $S(b) = \{\text{happy}, \text{run}, ...\}$; $P(b) = \{\text{children}...\}$; P'(b) = nouns.

a = run; $S(a) = \{children, in, ...\}$; $P(a) = \{run, have run, had run...\}$; P'(a) = verbs.

 $R(a) = \{in, the, garden, and, have fun...\}; R'(a)$ is a set of words which are equal to a, i.e. verbs, so by theorem 2: R'(a) = P'(a).

Condition of homogeneity: $S(a) \cap P(b) = \{children\}, S(b) \cap P(a) = \{run\}.$

 $K(a) = \{$ **children, run\}** is verbs (by the condition of equality), so by theorem 3: K'(a) = P'(a).

Class K(a) and division from mixed cells R(a) are equal as they contain verbs, so by theorem 4: K(a) = R(a).

Semantic role labelling

Semantic Role Labeling

Event - ?

- XYZ corporation bought the stock.
- They sold the stock to XYZ corporation.
- The stock was bought by XYZ corporation.
- The purchase of the stock by XYZ corporation...
- The stock purchase by XYZ corporation...

Participants - ?

Roles - ?

Semantic Role Labeling

Semantic roles are representations that express the abstract role that arguments of a predicate can take in the event; these can be very specific, like the *BUYER*, abstract like the *AGENT*, or super-abstract (the *PROTO-AGENT*). These roles can both represent general semantic properties of the arguments and also express their likely relationship to the syntactic role of the argument in the sentence. *AGENTS* tend to be the subject of an active sentence, *THEMES* the direct object, and so on.

Semantic role labeling is the task of assigning roles to the constituents or phrases in sentences.

Selectional restrictions, the semantic sortal restrictions or preferences that each individual predicate can express about its potential arguments, such as the fact that the theme of the verb eat is generally something edible.

Agent: a participant which the meaning of the verb specifies as doing or causing something, possibly intentionally. Examples: subjects of *kill*, *eat*, *hit*, *smash*, *kick*, *watch*.

Causer: inanimate or non-volitional entity which is directly involved in the causation of an event. *The rain ruined the crop*.

Patient: a participant which the verb characterizes as having something happen to it, and as being affected by what happens to it. Examples: objects of *kill*, *eat*, *hear*, *love*.

Experiencer: a participant who is characterized as aware of something. Examples: subject of *love*, object of *annoy*.

Theme: a participant which is characterized as changing its position or condition, or as being in a state or position. Examples: objects of *give*, *hand*, subjects of *walk*, *die*.

Location: the thematic role associated with the NP expressing the location in a sentence with a verb of location. Examples: *They stayed at the village*.

Source: object from which motion proceeds. Examples: *Jennifer walked home from the office*.

Goal: object to which motion proceeds. Examples: Jennifer walked <u>home</u> from the office.

Stimulus: whatever causes a psychological response (i.e. Positive or negative) in the experiencer. *The situation scares me*.

Content: the content of a psychological state or of a representation. It expresses an idea or a mental representation that is entertained or perceived. *The jury heard his testimony*.

Instrument: inanimate entity used by an agent or experiencer in order to do something to a patient or theme. *John opened the lock with a key*.

Recipient: animate entity which receives or acquires something. *John gave Mary a book.*

- She ran [from the post-office]_{SOURCE} [via the railway station]_{PATH} [to the bus-stop.]_{GOAL}
- Jenny walked [from school]_{PATH}
- Walter drove [home]_{GOAL}
- The coin rolled [across the floor]_{PATH}

- (22.3) *John broke the window*. AGENT THEME
- (22.4) John broke the window with a rock.

 AGENT THEME INSTRUMENT
- (22.5) The rock broke the window.

 INSTRUMENT THEME
- (22.6) *The window broke*. THEME
- (22.7) The window was broken by John.

 THEME AGENT

The set of thematic role arguments taken by a verb is thematic grid often called the **thematic grid**, θ -grid, or case frame.

- 1. John gave Mary a bouquet of roses;
- John baked Mary a chocolate cake;
- 3. The key opened the lock;
- 4. S. Holmes heard a piercing scream;
- 5. John enjoyed the novel;
- 6. We put the box on the shelf;
- 7. Mike drove to Jackie's house;
- 8. The burglar entered through the window.

Frame Semantics

Frame semantics is a theory of linguistic meaning developed by Charles J. Fillmore. It relates linguistic semantics to encyclopaedic knowledge.

Basic idea: one cannot understand the meaning of a single word without access to all the essential knowledge that relates to that word.

A word activates a frame of semantic knowledge relating to the specific concept it refers to (or highlights).

Frame Semantics

A **semantic frame** is a collection of facts that specify characteristic features, attributes, and functions of a denotatum, and its characteristic interactions with things necessarily or typically associated with it.

A **semantic frame** is a coherent structure of related concepts that are related such that without knowledge of all of them, one doesn't have complete knowledge of any one.

Words not only highlight individual concepts, but also specify a certain perspective from which the frame is viewed.

Frame Semantics

BUYER	buy	GOODS	(SELLER)	(PRICE)
subject		object	from	for
Alfred	bought	the book	from Olivia	for 10 dollars
Alfred	bought	them		for 1 dollar
Alfred	bought	a bicycle	from Sarah	

Semantic Roles: Problems

- a. The cook opened the jar with the new gadget.
- b. The new gadget opened the jar.
- a. Shelly ate the sliced banana with a fork.
- b. *The fork ate the sliced banana.

The Proposition Bank

https://propbank.github.io/

The Proposition Bank, generally referred to as PropBank, is a resource of sentences annotated with semantic roles.

The Proposition Bank

(22.11) agree.01

Arg0: Agreer Arg1: Proposition

Arg2: Other entity agreeing

Ex1: [Arg0 The group] agreed [Arg1 it wouldn't make an offer].

Ex2: [ArgM-TMP Usually] [Arg0 John] agrees [Arg2 with Mary]

[Arg1 on everything].

(22.12) **fall.01**

Arg1: Logical subject, patient, thing falling

Arg2: Extent, amount fallen

Arg3: start point

Arg4: end point, end state of arg1

Ex1: [Arg1 Sales] fell [Arg4 to \$25 million] [Arg3 from \$27 million].

Ex2: [Arg1 The average junk bond] fell [Arg2 by 4.2%].

The Proposition Bank: ArgMs, (ArgMTMP, ArgM-LOC, etc)

TMP when? yesterday evening, now

LOC where? at the museum, in San Francisco

DIR where to/from? down, to Bangkok

MNR how? clearly, with much enthusiasm

PRP/CAU why? because ..., in response to the ruling

REC themselves, each other

ADV miscellaneous

PRD secondary predication ...ate the meat raw

- (22.17) [$_{Arg1}$ The price of bananas] increased [$_{Arg2}$ 5%].
- (22.18) [$_{Arg1}$ The price of bananas] rose [$_{Arg2}$ 5%].
 - (22.19) There has been a [Arg2 5%] rise [Arg1] in the price of bananas].

We'd like a system to recognize that *the price of bananas* is what went up, and that 5% is the amount it went up, no matter whether the 5% appears as the object of the verb *increased* or as a nominal modifier of the noun *rise*.

FrameNet

https://framenet.icsi.berkeley.edu/fndrupal/

A **frame** in FrameNet is a background knowledge structure that defines a set of frame-specific semantic roles, called **frame elements**, and includes a set of predicates that use these roles. Each word evokes a frame and profiles some aspect of the frame and its elements. The FrameNet dataset includes a set of frames and frame elements, the lexical units associated with each frame, and a set of labeled example sentences.

For example, the *change_position_on_a_scale* frame is defined as follows:

This frame consists of words that indicate the change of an Item's position on a scale (the Attribute) from a starting point (Initial value) to an end point (Final value).

FrameNet: Roles

Core Roles						
ATTRIBUTE	The ATTRIBUTE is a scalar property that the ITEM possesses.					
DIFFERENCE	The distance by which an ITEM changes its position on the scale.					
FINAL_STATE	A description that presents the ITEM's state after the change in the ATTRIBUTE's					
	value as an independent predication.					
FINAL_VALUE	The position on the scale where the ITEM ends up.					
INITIAL_STATE	A description that presents the ITEM's state before the change in the AT-					
	TRIBUTE's value as an independent predication.					
INITIAL_VALUE	The initial position on the scale from which the ITEM moves away.					
ITEM	The entity that has a position on the scale.					
VALUE_RANGE	A portion of the scale, typically identified by its end points, along which the					
	values of the ATTRIBUTE fluctuate.					
Some Non-Core Roles						
DURATION	The length of time over which the change takes place.					
SPEED	The rate of change of the VALUE.					
GROUP	The GROUP in which an ITEM changes the value of an					
D: 44.2	ATTRIBUTE in a specified way.					

Figure 22.3 The frame elements in the change_position_on_a_scale frame from the FrameNet Labelers Guide (Ruppenhofer et al., 2006).

FrameNet: Roles

- (22.20) [ITEM Oil] rose [ATTRIBUTE in price] [DIFFERENCE by 2%].
- (22.21) [ITEM It] has increased [FINAL STATE to having them 1 day a month].
- (22.22) [ITEM Microsoft shares] fell [FINAL_VALUE to 7 5/8].
- (22.23) [ITEM Colon cancer incidence] *fell* [DIFFERENCE by 50%] [GROUP among men].
- (22.24) a steady *increase* [INITIAL_VALUE from 9.5] [FINAL_VALUE to 14.3] [ITEM in dividends]
- (22.25) a [DIFFERENCE 5%] [ITEM dividend] increase...

All words for this frame



advance climb decline decrease diminish dip	explode fall	mushroom plummet reach rise rocket shift	swing triple tumble NOUNS:	explosion fall fluctuation gain growth hike increase	shift tumble ADVERBS: increasingly
------------------------------------------------------------	-----------------	---------------------------------------------------------	-----------------------------	------------------------------------------------------------------------	-------------------------------------------

Semantic Role Labeling

```
function SEMANTICROLELABEL(words) returns labeled tree

parse ← PARSE(words)
for each predicate in parse do
    for each node in parse do
        featurevector ← EXTRACTFEATURES(node, predicate, parse)
        CLASSIFYNODE(node, featurevector, parse)
```

Figure 22.4 A generic semantic-role-labeling algorithm. CLASSIFYNODE is a 1-of-*N* classifier that assigns a semantic role (or NONE for non-role constituents), trained on labeled data such as FrameNet or PropBank.

Semantic Role Labeling

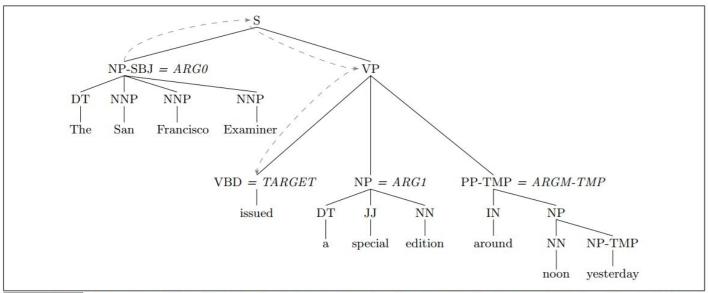


Figure 22.5 Parse tree for a PropBank sentence, showing the PropBank argument labels. The dotted line shows the **path** feature NP \uparrow S \downarrow VP \downarrow VBD for ARG0, the NP-SBJ constituent *The San Francisco Examiner*.

Features for Semantic Role Labeling

- The governing **predicate**: the predicate is a crucial feature since labels are defined only with respect to a particular predicate;
- The **phrase type** of the constituent: some semantic roles tend to appear as NPs, others as S or PP, and so on;
- The headword of the constituent;
- The headword part of speech of the constituent;
- The **voice of the clause** in which the constituent appears: passive sentences tend to have strongly different linkings of semantic roles to surface form than do active ones;
- The binary linear position of the constituent with respect to the predicate, either before or after;
- The named entity type of the constituent;
- The first words and the last word of the constituent;

A Neural Algorithm for Semantic Role Labeling

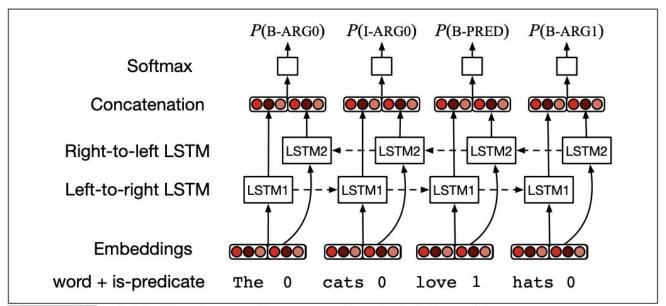


Figure 18.6 A bi-LSTM approach to semantic role labeling. Most actual networks are much deeper than shown in this figure; 3 to 4 bi-LSTM layers (6 to 8 total LSTMs) are common. The input is a concatenation of an embedding for the input word and an embedding of a binary variable which is 1 for the predicate to 0 for all other words. After He et al. (2017).

Evaluation of Semantic Role Labeling

Metrics:

- Precision;
- Recall;
- F-measure

Data sets:

- CoNLL-2005;
- CoNLL-2012

Selectional Restrictions

A selectional restriction is a semantic type constraint that a verb imposes on the kind of concepts that are allowed to fill its argument roles.

I want to eat someplace nearby.

Selectional restrictions are associated with senses, not entire lexemes:

The restaurant **serves** green-lipped mussels.

Which airlines **serve** Denver?

Selectional Restrictions

In rehearsal, I often ask the musicians to imagine a tennis game.

Radon is an odorless gas that can't be detected by human senses.

To diagonalize a matrix is to find its eigenvalues.

Some links

Semantic Role Labeling Demo: http://cogcomp.org/page/demo_view/srl

Semantic Role Labeler: http://barbar.cs.lth.se:8081/

Thank you for your attention!