CORRECTING COLLOCATION ERRORS IN LEARNERS' WRITING BASED ON PROBABILITY OF SYNTACTIC LINKS

Azimov Alexander

Bolshakova Elena

CMC MSU

Collocation errors

- Lexical errors

Violate norms of lexical combinability;

- Grammatical errors

Coordination mistakes, etc.

Why do they occur?

Collocation errors:

- Unusual for native speakers;
- Typical errors in English Second Learners (ESL) writing.

Strategy of word-by-word translation is core of the problem

	Word-by-word translation
Красивый мужчина	Beautiful man
Сильный дождь	Powerful rain
Великий художник	Great painter

Why do they occur?

Collocation errors:

- Unusual for native speakers;
- Typical errors in ESL writing.

Strategy of word-by-word translation is core of the problem

	Word-by-word translation	Proper correction
Красивый мужчина	Beautiful man	Handsome man
Сильный дождь	Powerful rain	Heavy rain
Великий художник	Great painter	Great artist

Correctors

- Orthographic errors
- Some types of syntax errors

There is no currently available corrector able to detect collocation errors

Correcting collocation errors using probability theory

2008:

The problem of error correction within a sentence *S* considered as the task to find most probable correcting sentence ***V****, among possible sentences *V*, given sentence *S*

$$V^* = arg \max_V \{P(V|S)\}$$

Using the Web to Automatically Correct Lexico-Syntactic Errors

2008 г. M. Hermet , A. Désilets, S. Szpakowicz

Main features	Description
Language	French
Types of correcting errors	Articles, prepositions
Collocation extraction	Full parsing
Substitute words	Database made by expert
Correcting algorithm	Frequency statistics from Yahoo

Automatic Collocation Suggestion in Academic Writing

2010, J. Wu, Y. Chang, T. Mitamura, J. S. Chang

Main features	Description
Language	English
Types of correcting errors	Nouns, verbs
Collocation extraction	Full parsing + N-gramms
Substitute words	The set of substitutes words matches with considered words
Correcting algorithm	Maximum entropy classifier

A Web-based English Proofing System for English as a Second Language Users

2009, X. Yi, J. Gao, W. B. Dolan

Main features	Description
Language	English
Types of correcting errors	Verb-Noun; Verb-Preposition-Noun
Collocation extraction	Syntactic templates
Substitute words	--
Correcting algorithm	Frequency analysis of snippets from BING search engine

Current state of error correction

- Particular methods for parts of speech;
- No general method
- No use of native language model of the writer;
- Experts made databases of correcting substitute words;
- The use of raw frequency from search engine.

Our plans…

Key ideas:

- No dependence on part of speech;
- Automatic generation possible substitute words and correcting paraphrases;
- Detection and correction several errors in one sentence.
- No dependence on language;

Correcting collocation errors using probability theory

The problem of error correction within a sentence *S* may be considered as the task to find most probable correcting sentence ***V****, among possible sentences *V*, given sentence *S*

$$V^* = arg\max_V\{P(V|S)\} = arg\max_V\{P(S|V)P(V)\}$$

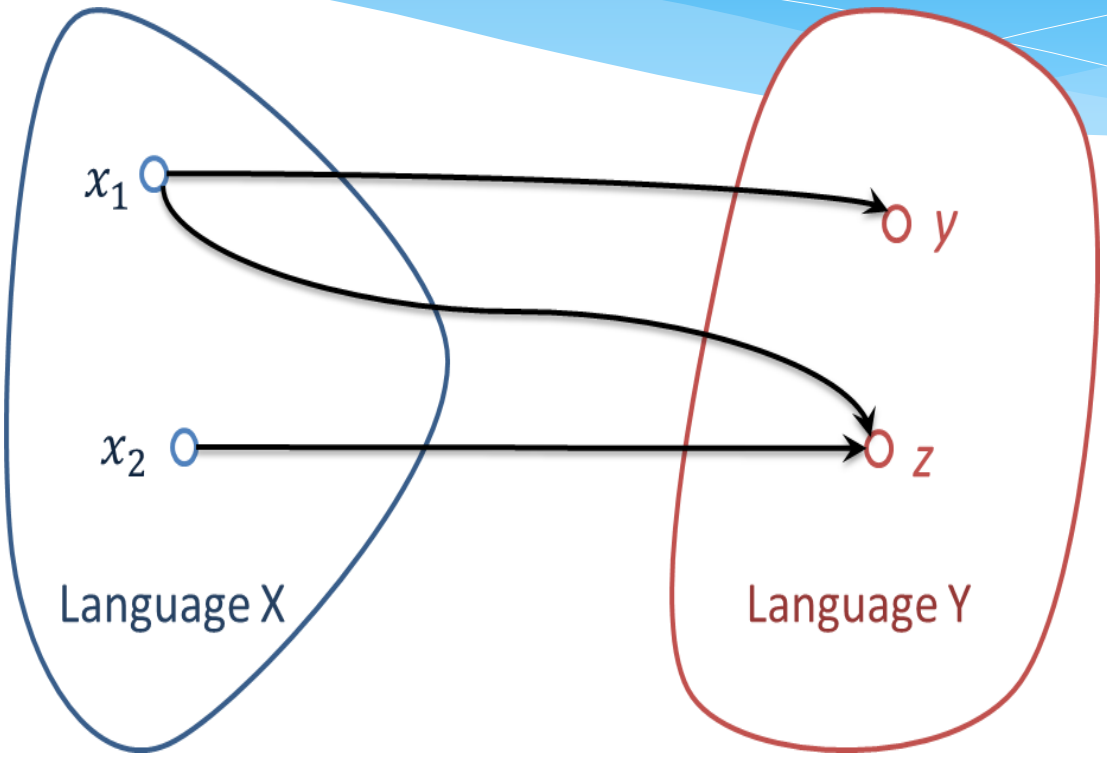
1. Determine probability of sentence *V* (paraphrase) as correcting variant for *S*;
2. Determine probability of sentence *V*.

Main assumptions

- Collocation errors don't change the syntactic structure of sentences ;
- Independence of collocation errors in the sentence.

To determine the probability of paraphrases we need to determine the probability of their components, i.e. substitute words.

Substitute words: Map *Translate*



A set of ordered pairs $\langle x, y \rangle$ where

word *x* belongs to the source language *X* and has a set of translation equivalents {*y*} from the target language *Y*.

Substitute words: generation

A wrong word , as well as a correct one, both are images of certain word *x* from *X*:

$$DoubleTranslation(y) = \{z \in Y \mid \exists x \in X : z \in Translate(x) \& y \in Translate(x)\}$$

To reduce the set we take into account only synonyms of word *y*:

$$Substitutes(y) = \{z \in Y \mid z \in DoubleTranslation(y) \ \& \ z \in Synonyms(y)\}$$

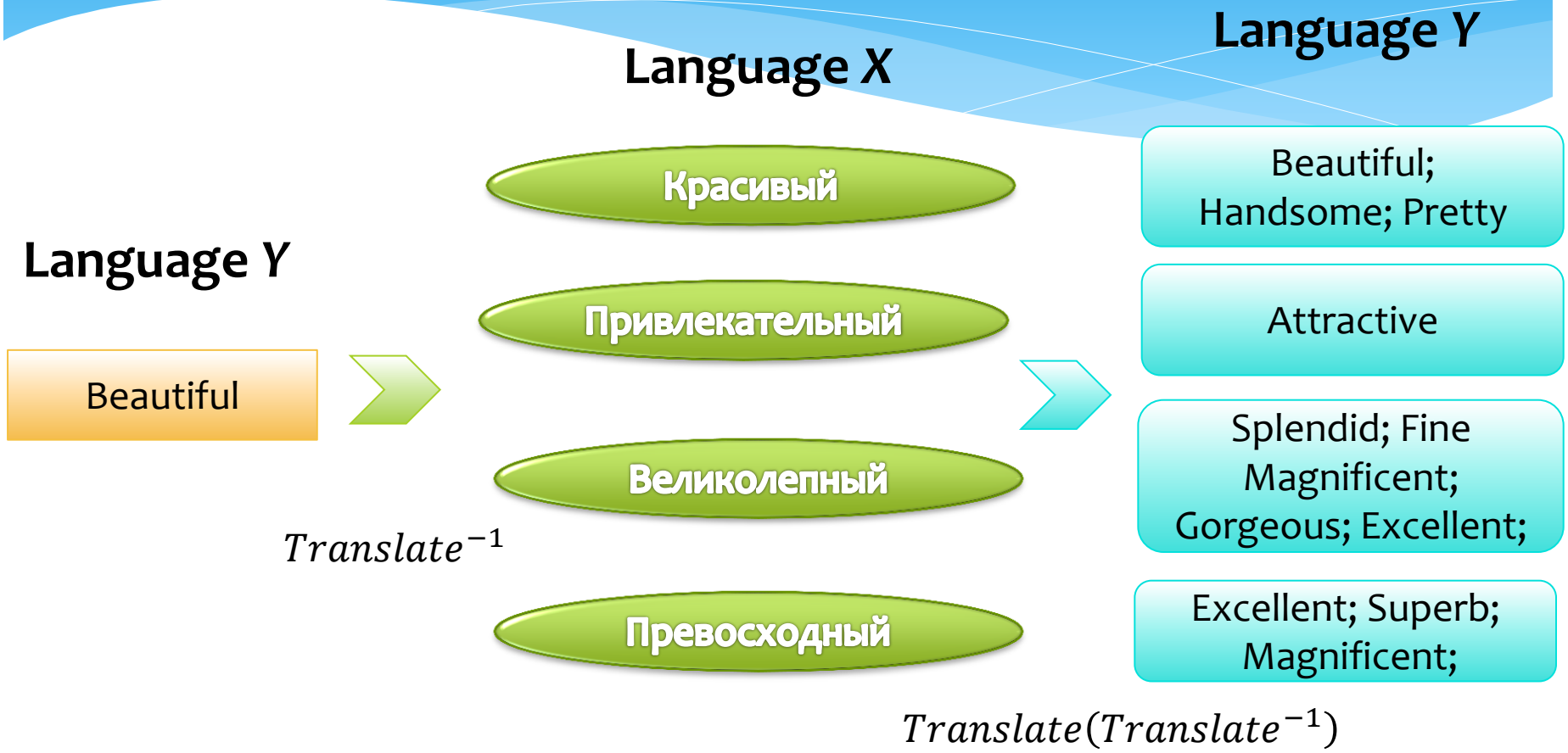
Substitute words: an example

Language *Y*

Beautiful

Substitute words: an example

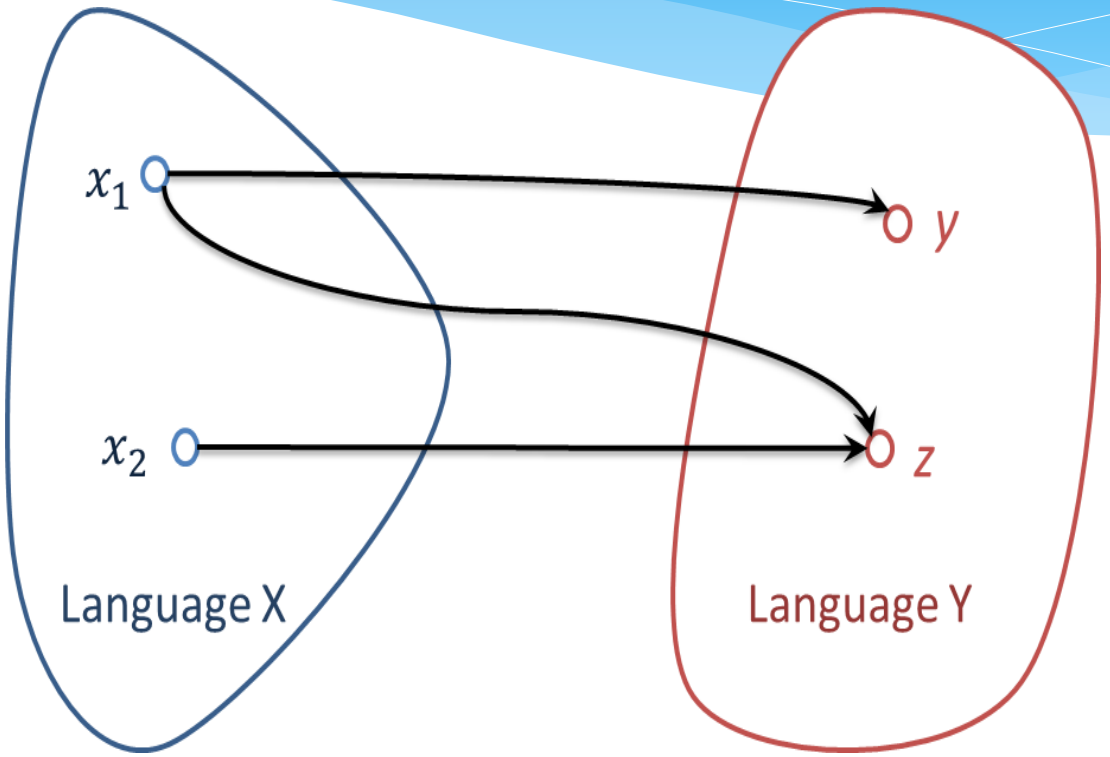
Substitute words: an example



Set of substitute words for word «beautiful»:

attractive, fine, gorgeous, handsome, pretty

Substitute words: ranking



$p_+(y|x)$ -- conditional probability that word x is preimage of word y;

$p_-(x|y)$ -- conditional probability that word y is preimage of word x.

Substitute words: probability

The conditional probability of substitute word *z*, given words *y* and *x*, and *x* is a preimage of *y*:

$$p(z|y,x) = p_{-}(x|y)p_{+}(z|x)$$

Conditional probability of substitute word *z* for a given word *y*:

$$p_{dt}(z|y) = \sum_{\{x|y \in Translation(x) \& z \in Translation(x)\}} p_{-}(x|y)p_{+}(x|y)$$

Paraphrase probability

The paraphrase probability equals to product of probabilities of its substitute words:

$$p(S|V) = \prod_i p_{dt}(s_i|v_i)$$

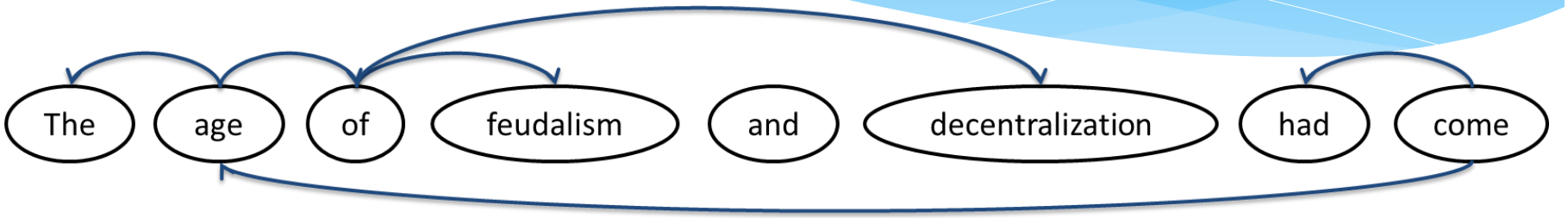
Correcting collocation errors using probability theory

The problem of error correction within a sentence *S* may be considered as the task to find most probable correcting sentence ***V****, among possible sentences *V*, given sentence *S*

$$V^* = arg\max_V\{P(V|S)\} = arg\max_V\{P(S|V)P(V)\}$$

1. Determine probability of sentence *V* as correcting variant for *S*;
2. Determine probability of sentence *V*.

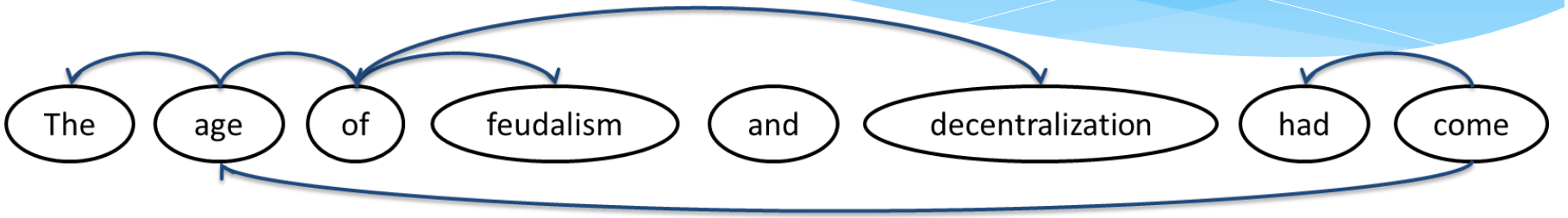
Dependency tree of sentence



The word v_1 is *ancestor* of v_2, if directed path from vertex v_1 to vertex v_2 exists.

Let *ancestors*(v_i) denote the set of all *ancestors* for word v_i .

Ancestor set: an example



Word	Ancestors
The	age, come
age	come
of	age, come
feudalism	of, age, come
decentralization	of, age, come
had	come
come	--

Sentence probability

We assume conditional independence of each word v_i from all other words except its *ancestors*.

So we computed the joint probability of the words from the sentence, given the particular sentence parse tree

$$\mathrm{P}(V) = p(v_1, \ldots, v_n) = \prod_{i=1}^{n} p(v_i | ancestors(v_i))$$

Probabilities computation

We use word syntactic link statistics gathered on some text collection:

$$p(v_i|parents(v_i)) = \frac{N(v_i, ancestors(v_i))}{N\big(ancestors(v_i)\big)}$$

where $N(v_i, ancestors(v_i))$ and $N(ancestors(v_i))$ are frequencies of corresponding syntactically related group of words

Correcting paraphrase

$$V^* = arg\max_V \{P(S|V)P(V)\}$$

$$V^* = arg\max_{v_1,\ldots,v_k} \prod_{i=1}^{k} (p_{dt}(s_i|v_i)p(v_i|ancestors(v_i))),$$

where $s_i \in Substitutes(v_i)$

Degree function

Implementation for experiments

1. Collocation errors in English writing;
2. Correcting only one error in each sentence;
3. For each vertex of parse tree only two *ancestors* are considered.

Database for experiments

Substitute words database:

- We have built substitute sets for more then 29 thousand of words.

Syntactic links database - Stanford Parser was used:

- 220 billion of words were processed;
- Extracted 18 billion of syntactically linked word pairs;
- 65 billion of syntactically linked word triples.

Correcting procedure

1. Syntactic analysis to obtain dependency parse tree.
2. Generating for each word from *S* its *Substitutes* set and compute conditional probabilities.
3. Generating for each word from *S* a paraphrase *V based* on the generated *Substitutes* set of the word, thus forming a set of paraphrases for *S*.
4. Calculating the value of $Degree$ function for sentence *S* and its paraphrases.
5. If some paraphrases have *Degree* values that exceed the *Degree* value of *S*, signal a collocation error.
6. Building a ranked list of paraphrases with high $Degree$ values as candidate corrections for human editor.

Evaluation

Erroneous sentence	Proper correction
I think it is a **spend** of my money.	I think it is a **waste** of my money.
To make **understandable**.	To make **plain**.
I have **done** a mistake.	I have **made** a mistake.
The jar was full **with** oil.	The jar was full **of** oil
This is great **painter**.	This is great **artists**.
The **ghost** of the opera.	The **phantom** of the opera.

80% of collocation errors were detected.

87% of candidate corrections lists included the proper correction.

Mean reciprocal rank

$$MRR = \frac{1}{L}\sum \frac{1}{r}$$

where *L* is the number of sentences we used in our experiments

r is the rank of proper correction in a candidate corrections list.

Rank of proper correction	r = 1	r ≤ 2	r ≤ 3	r ≤ 100	MRR
k	35	45	48	49	0.5

Results

We proposed a novel method for collocation errors correction in learners' writing with the next main features:

- Correction of errors based on probabilities of word syntactic links
- Automatic generation and ranking of possible correcting paraphrases;
- No dependence on part of speech

The evaluation of the method showed promising results.

Future research

- Use of Bayesian networks to make the detecting procedure more efficient and test our method on sentences with several collocation errors;
- Expansion of *Substitutes* sets with word forms and homophones in order to detect additional type of collocation errors.

Questions?

mitradir@gmail.com