# Prototype of Semantic Model of Polish for Automatic Speech Recognition

Bartosz Ziółko

www.dsp.agh.edu.pl

rozpoznawaniemowy.blogspot.com

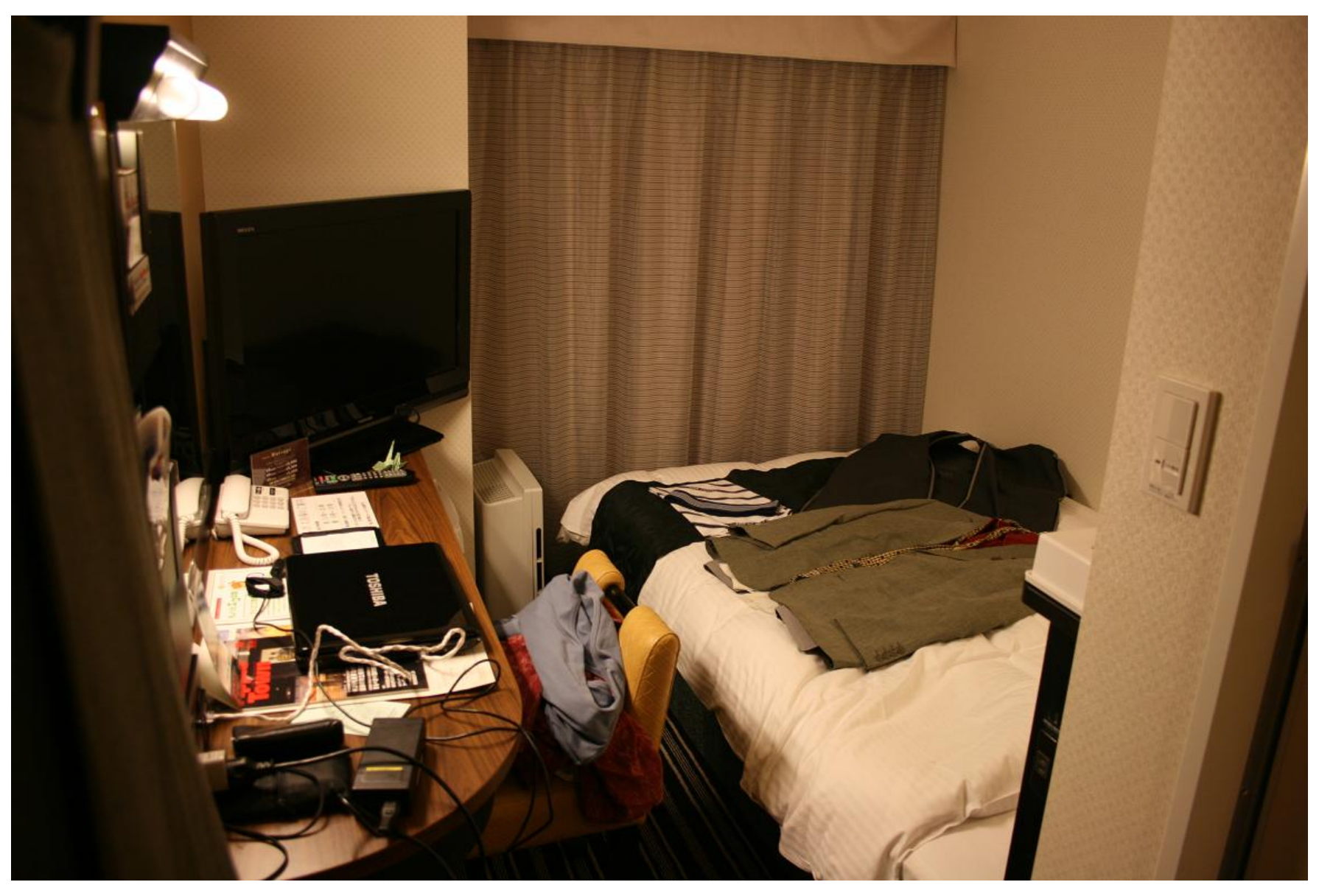
## Principle of compositionality

Typically, it is possible to get a meaning of word constructions like big dog from separate, independent meanings of big and dog. Semantics of natural language sentences and phrases can be composed from the semantics of their subparts.

*'Ce corps qui s'appelait et qui s'appelle encore le saint empire romain n'était en aucune manière ni saint, ni romain, ni empire.'* - Voltaire

*'This body, which called itself and still calls itself the Holy Roman Empire, was neither Holy, nor Roman, nor an Empire.'*

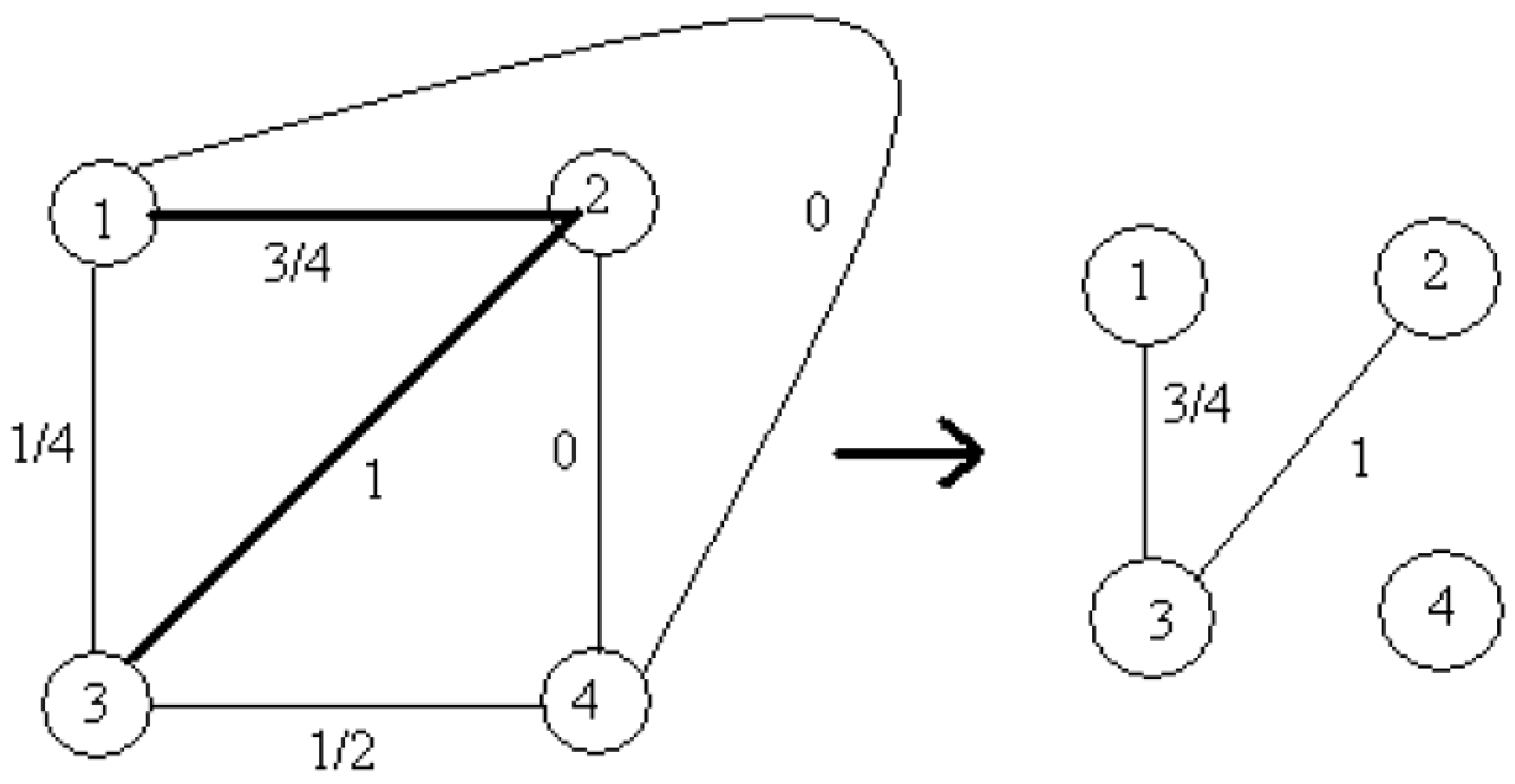
$$d_{ij} = \sum_{k=1}^{K} s_{ik} s_{jk} \qquad d'_{ij} = d_{ij} / \max_{i<j} \{d_{ij}\}$$



Business class in hotels is for people going to do business, but in case of airlines, it is a name for luxury which has little to do with business.

Training corpus: Big John has a house (1). Big John has a black, aggressive cat (2). The black aggressive cat has a small mouse (3). The small mouse is a mammal (4).

| topic | big | John | has | house | black | aggr. | cat | small | mouse | is | mammal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 3' | 7/8 | 7/8 | 15/8 | 1/2 | 11/8 | 11/8 | 11/8 | 1 | 1 | 0 | 0 |



Create an undirected, complete graph with topics as nodes and $d'_{ij}$ as weights of edges. Let us define path weight

$$p_{ij} = \prod_{(a,b) \in P(i,j)} d'_{ab}$$

| topic | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 4 | 3 | 1 | 0 |
| 2 | 3 | 6 | 4 | 0 |
| 3 | 1 | 4 | 6 | 2 |
| 4 | 0 | 0 | 2 | 4 |

where *P(i,j)* is a sequence of path edges from *i* to *j*. In the case of a single edge *i* to *j* path weight is $d'_{ij}$ . For multiple edges - a product of similarities of all edges on the path.

For each node, we need to find *n* nodes with highest path weights between the nodes and the given, analysed topic node. It will allow us to define a list *N* of semantically related topics which consists of the *n* nodes with their measures.

S is recalculated to include impact of similar topics. Smoothed word-topic relations are expressed by matrix

$$\mathbf{S}' = [s'_{ik}] \qquad s'_{ik} = s_{ik} + \alpha^{-1} \sum_{j \in N} p_{ij} s_{jk}$$

Matrix stores counts of words present in particular topics. They can be represented as probability-like count function (not normalised to [0,1]).

1)Find *n* single edge paths with the highest measures $d'_{ij}$.

2)Check if the two edges path *P(i,m)* starting from the node *i* with the highest $d'_{ij}$, which was found in the step above and going through *j* to any other edge *m*, has a better measure $p_{im}$ than the lowest of the *n* solutions found in the step above. If it does than replace the lowest one with *m* in the list of *n* similar topics.

3)Conduct the step above for all other single node paths from the list apart from the lowest, *n*th element.

4)If there are any non single edge paths *P(i,j)* on the list on position different then *n*th, go on like step 2. Check if after adding any other edge, a measure of path $p_{ij}$ is higher than a measure of the *n*th position. Than replace the previous path with a new, longer path with higher $p_{ij}$.

$$P_{sem} = \max_{i} \sum_{k \in W} s'_{ik}$$

POLSKA PLATFORMA
BEZPIECZEŃSTWA WEWNĘTRZNEGO

PPBW