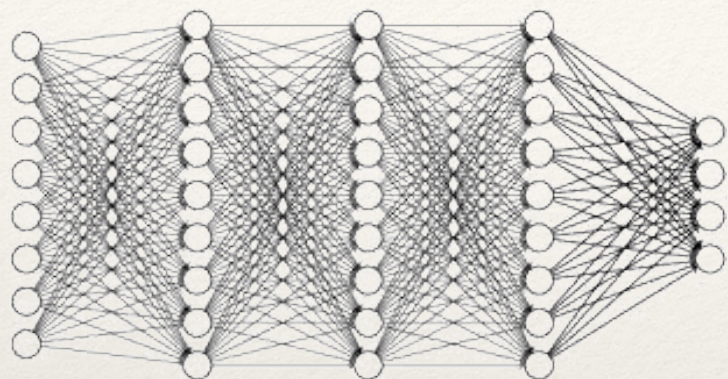

Sieci neuronowe do przetwarzania języka naturalnego - wybrane architektury i problemy

Paweł Morawiecki
IPI PAN

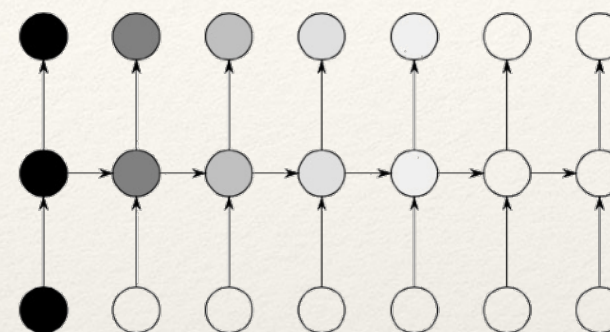
Sieci neuronowe stają się ważnym narzędziem dla NLP

- ❖ budowanie reprezentacji słów (word2vec)
- ❖ tematyczna klasyfikacja tekstów
- ❖ wykrywanie emocji i wydźwięku tekstu (ang. sentiment analysis)
- ❖ wykrywanie i grupowanie koreferencji
- ❖ coraz lepsze tłumaczenie tekstów
- ❖ Natural Language Processing with Deep Learning (Winter 2017), Stanford - dostępne na www.youtube.com

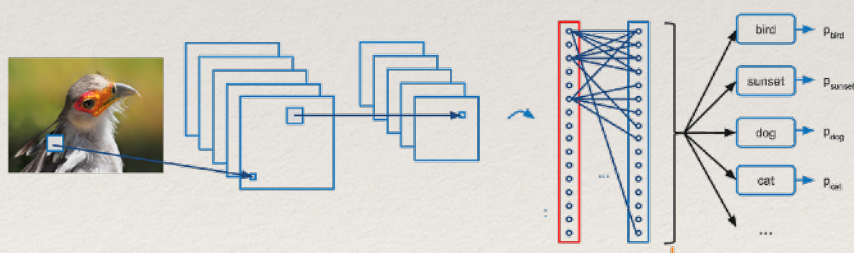
Architektury sieci można (umownie) podzielić na cztery grupy



ogólnego przeznaczenia
(fully connected)



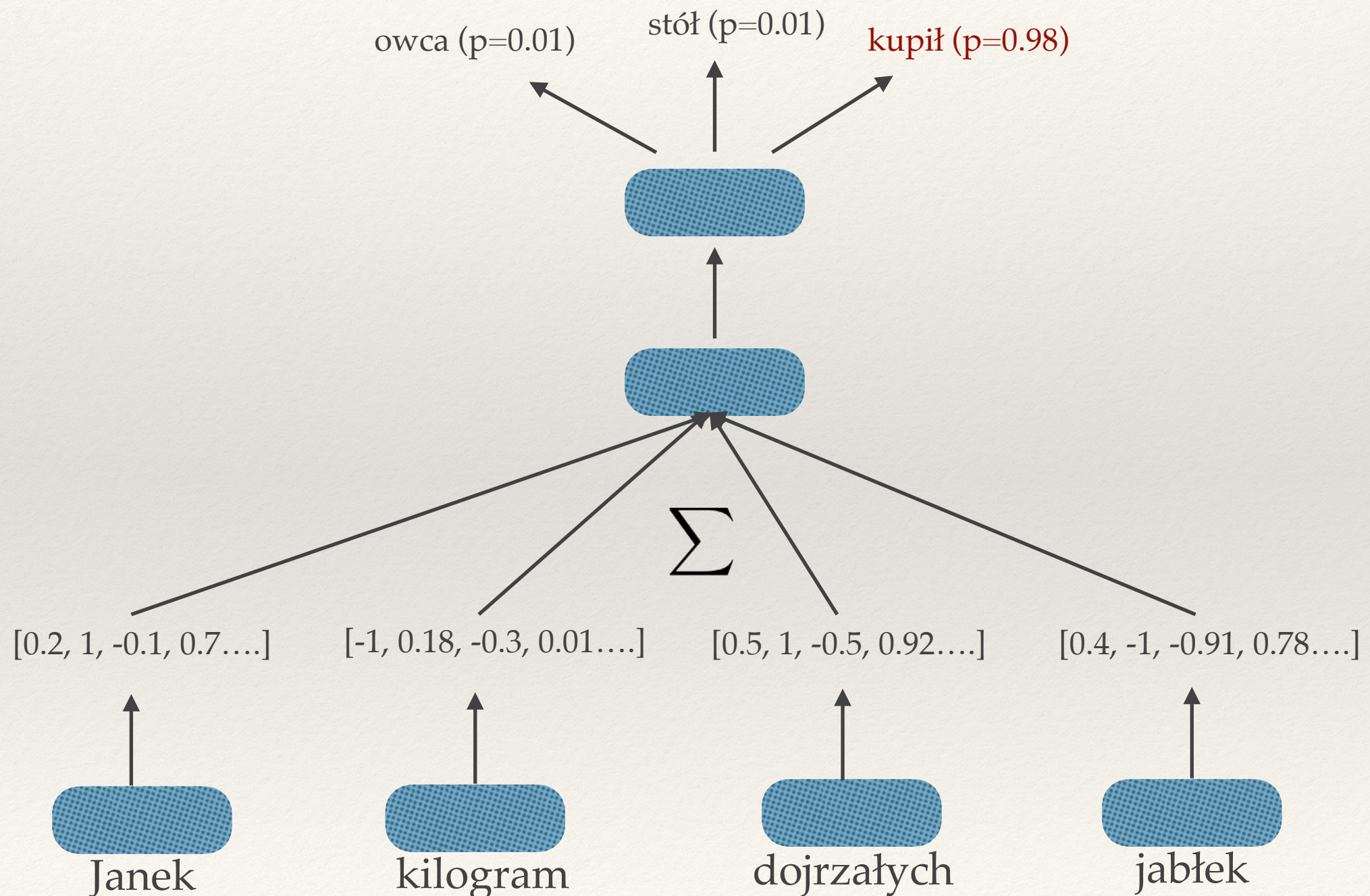
dla danych o charakterze sekwencyjnym
(RNN, LSTM, GRU...)



konwolucyjne
(głównie dla grafiki)

inne

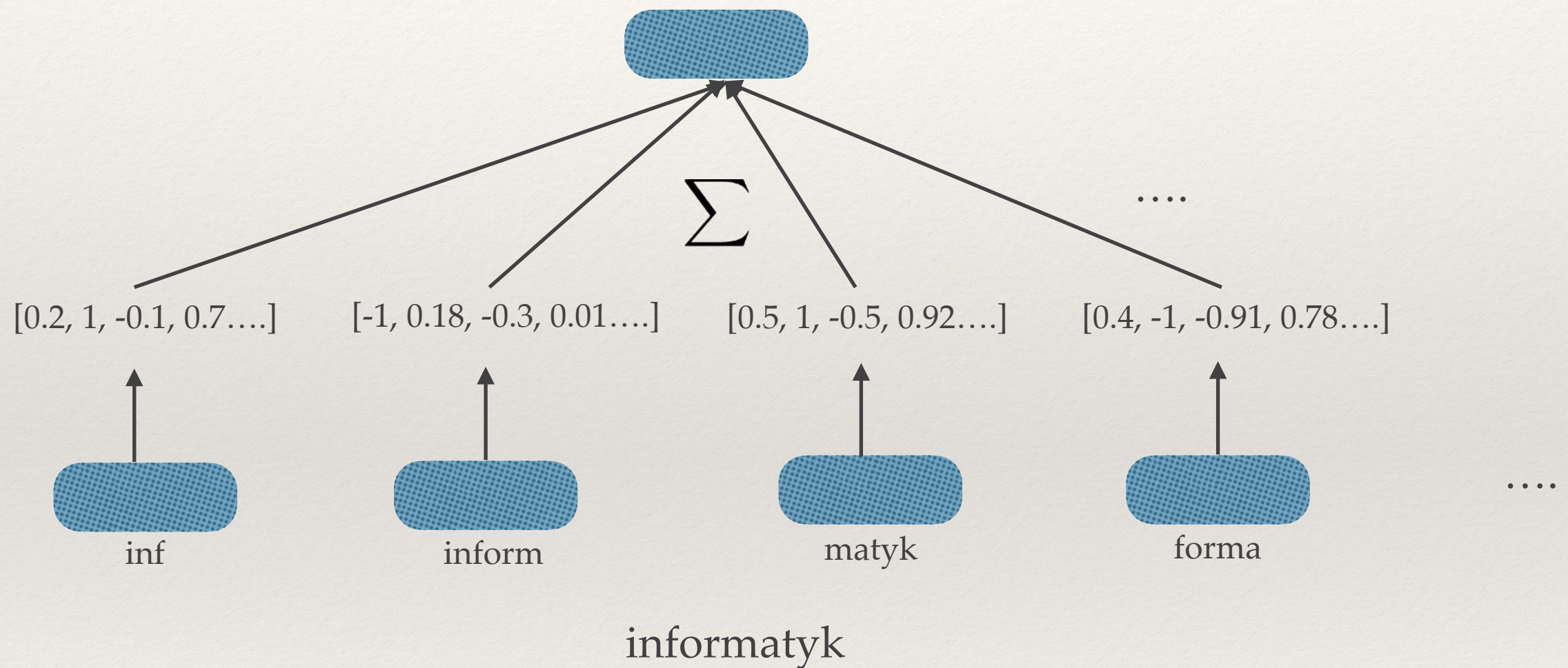
Sieć neuronowa pomaga znaleźć dobrą reprezentację słów



Nie ma jednego słusznego modelu

- ❖ model CBOW (Continuous Bag of Words) czyli słowo na podstawie kontekstu
- ❖ model skip-gram czyli kontekst na podstawie słowa

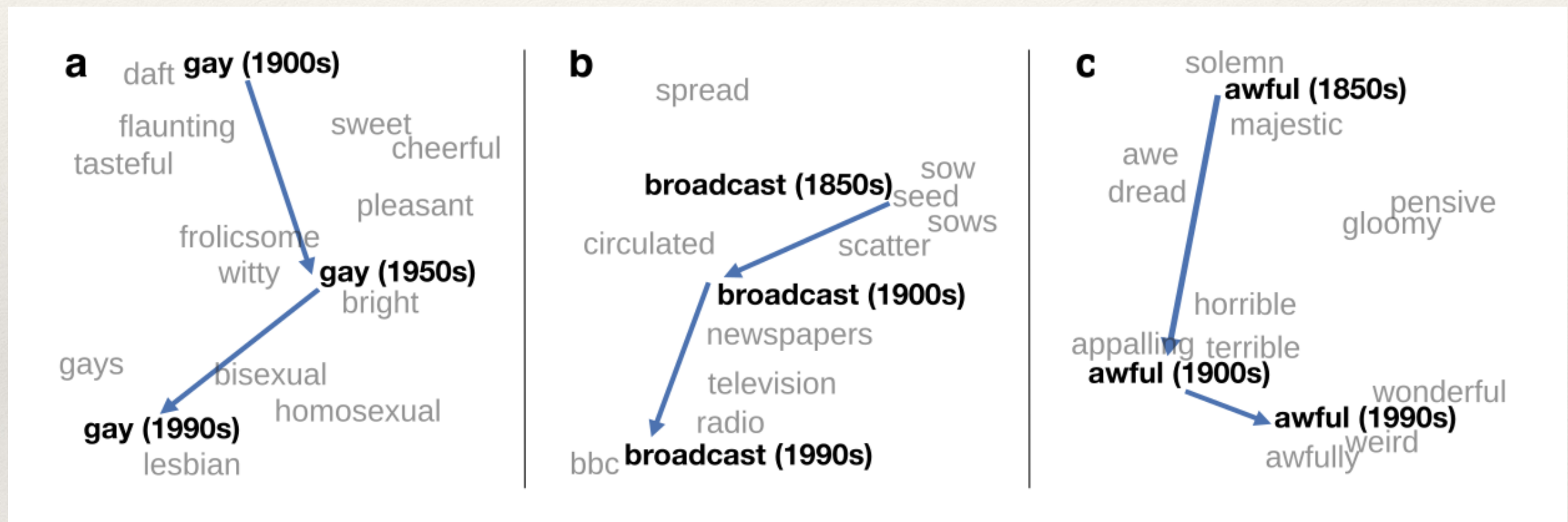
Warto budować reprezentację pod słów



Model z podsłowami lepiej radzi sobie z rzadkim słowami

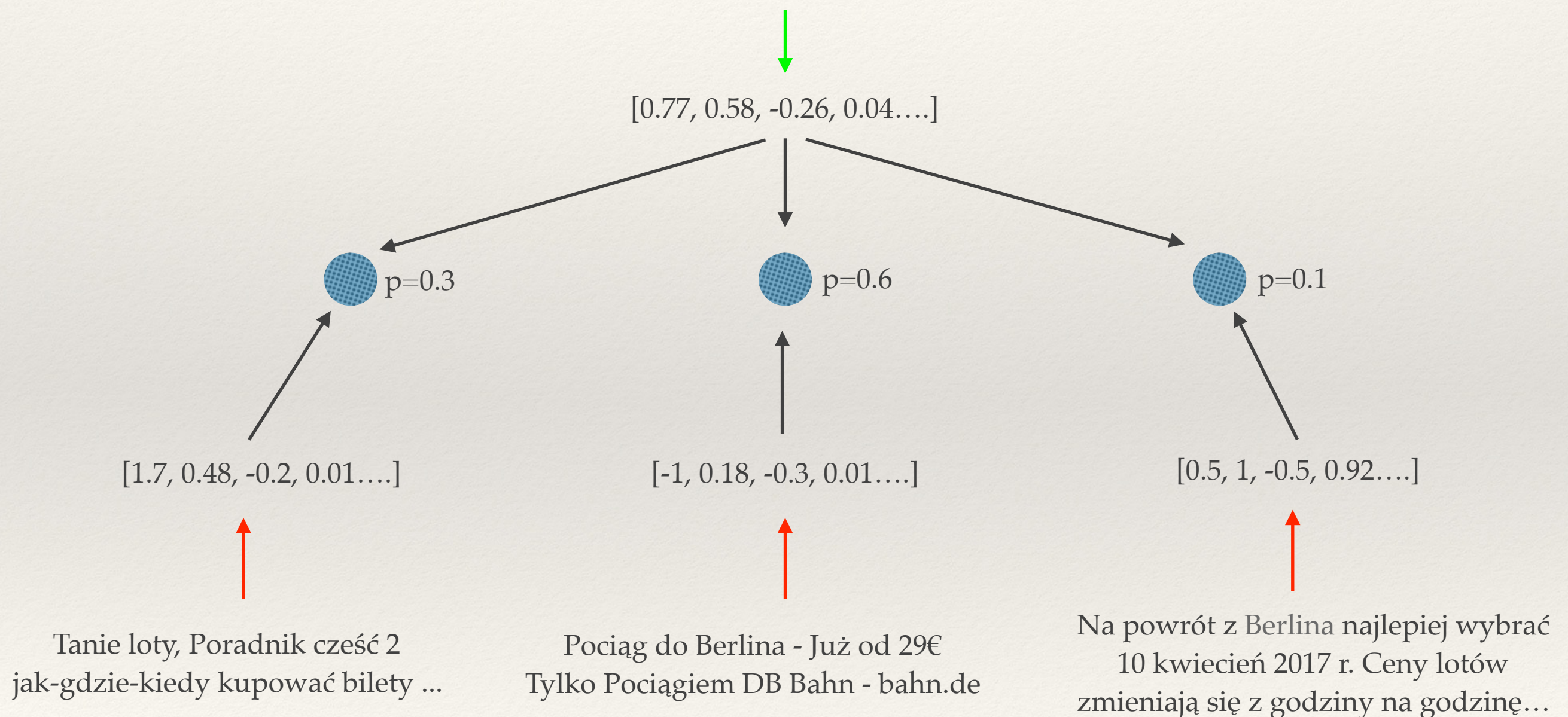
	tiling	tech-rich	english-born	dendritic
fastText (z reprezentacją podsłów)	tile flooring	tech-dominated tech-heavy	british-born polish-born	dendrite dendrites
word2vec	bookcases built-ins	technology-heavy .ixic	most-capped ex-scotland	epithelial p53

Reprezentacja typu „word2vec” znakomicie służy do śledzenia ewolucji znaczeniowej słów

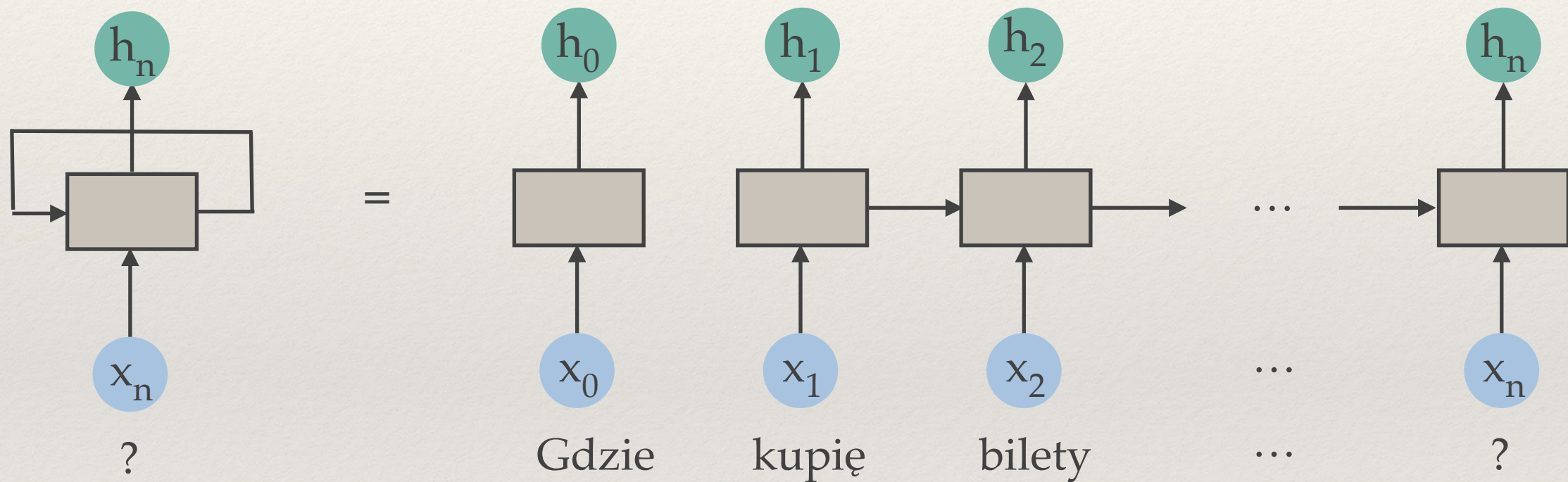


Reprezentacja całych zdań czy paragrafów też może być użyteczna

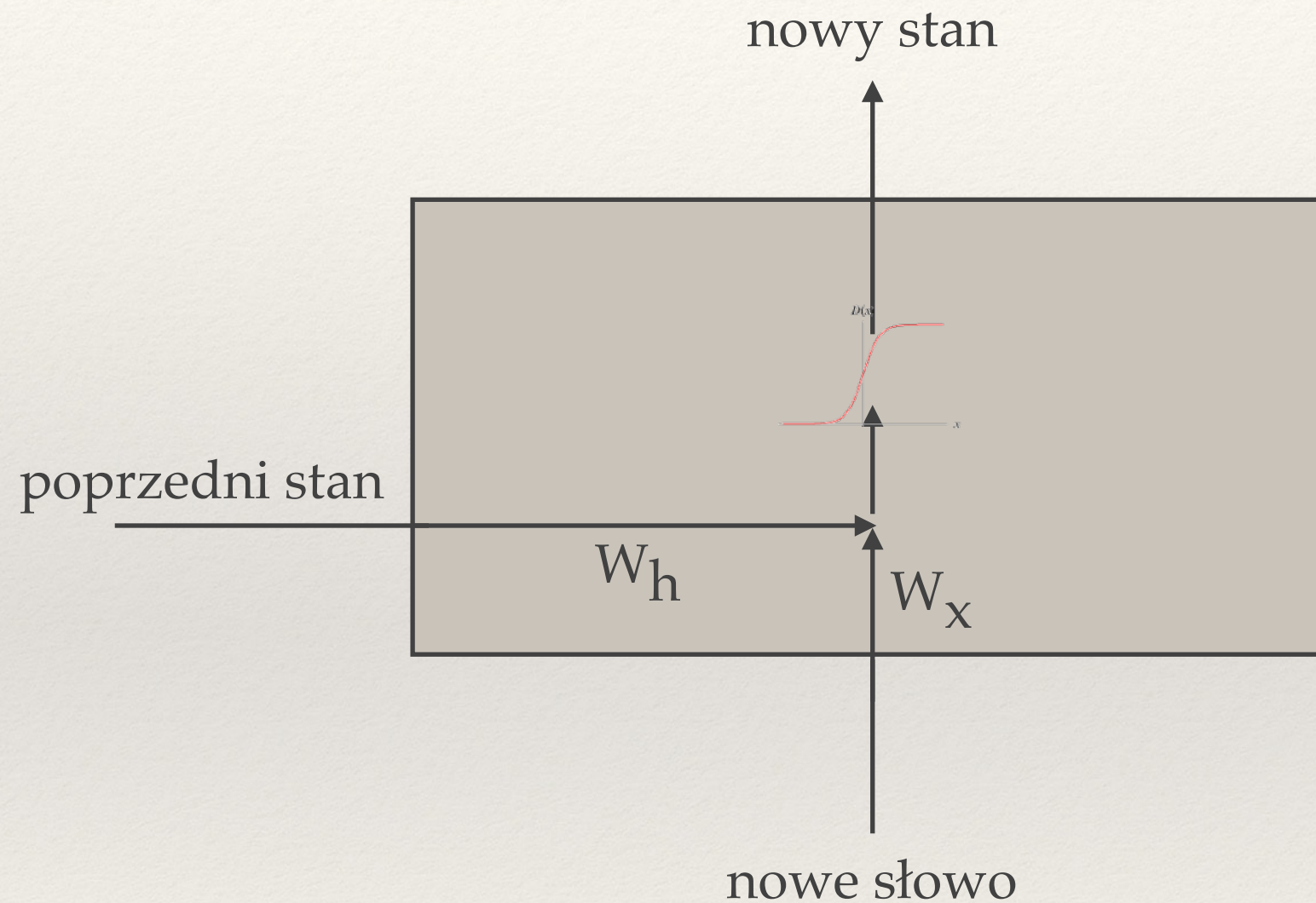
Gdzie najtańsze bilety do Berlina?



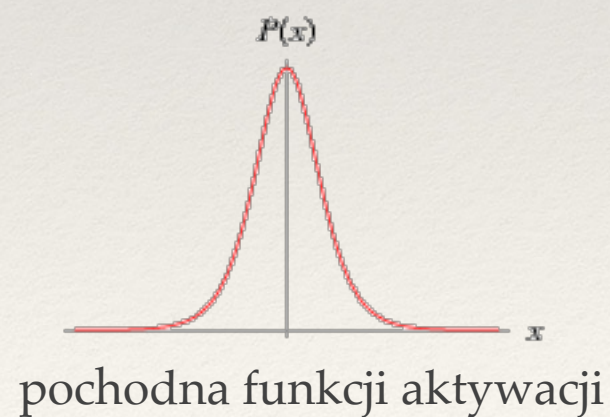
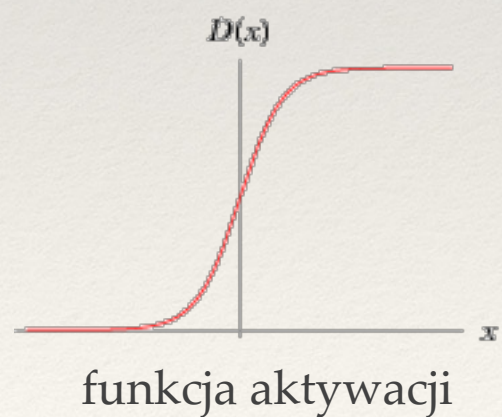
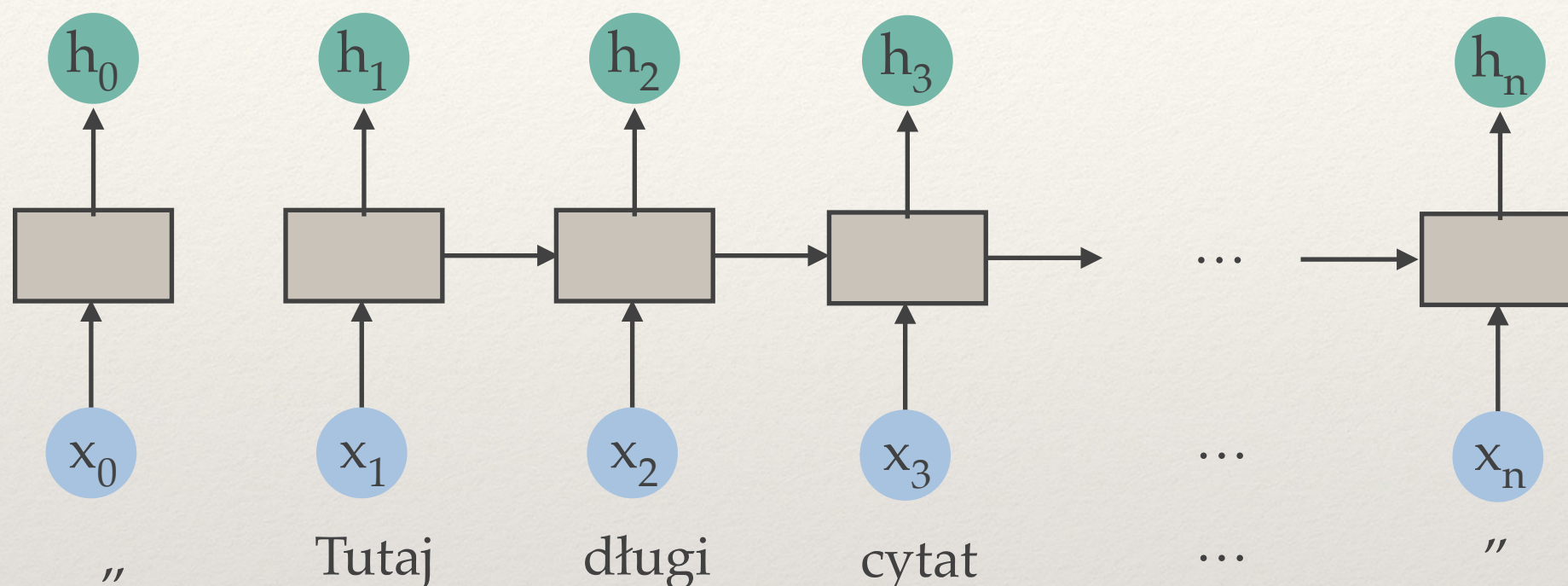
Sekwencyjne przetwarzanie tekstu wymaga bardziej elastycznej architektury sieci



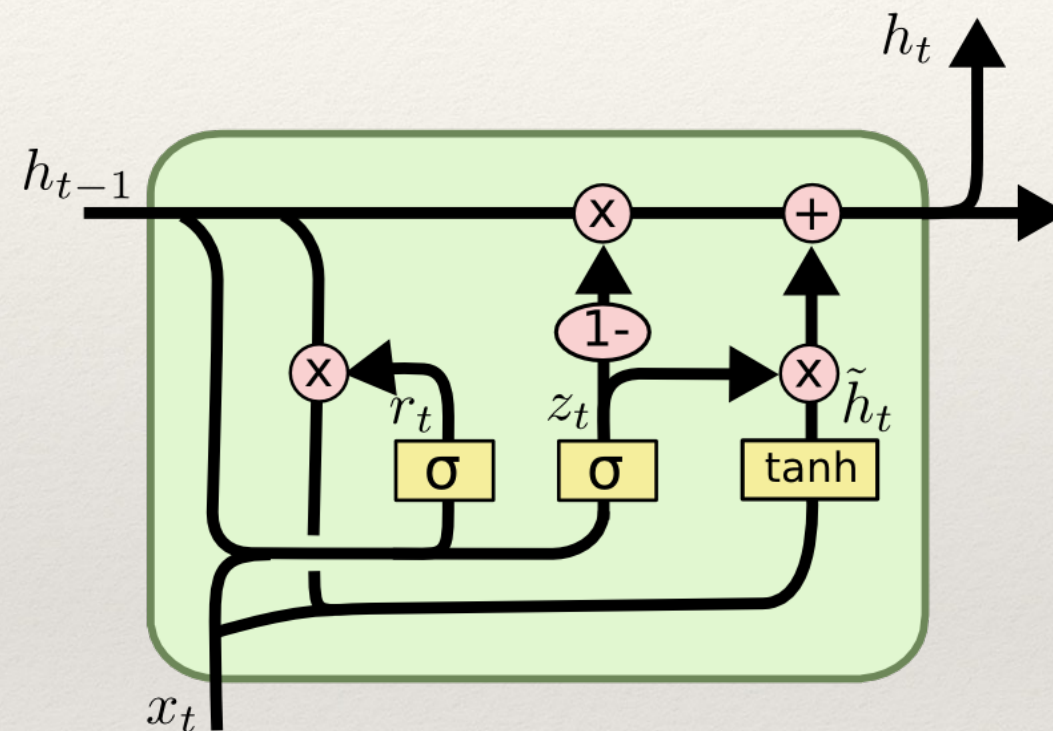
W sieci typu RNN stan wyjściowy zależy od wszystkich dotychczasowych informacji wejściowych



Podstawowy wariant RNN ma problem z wychwyceniem odległego kontekstu



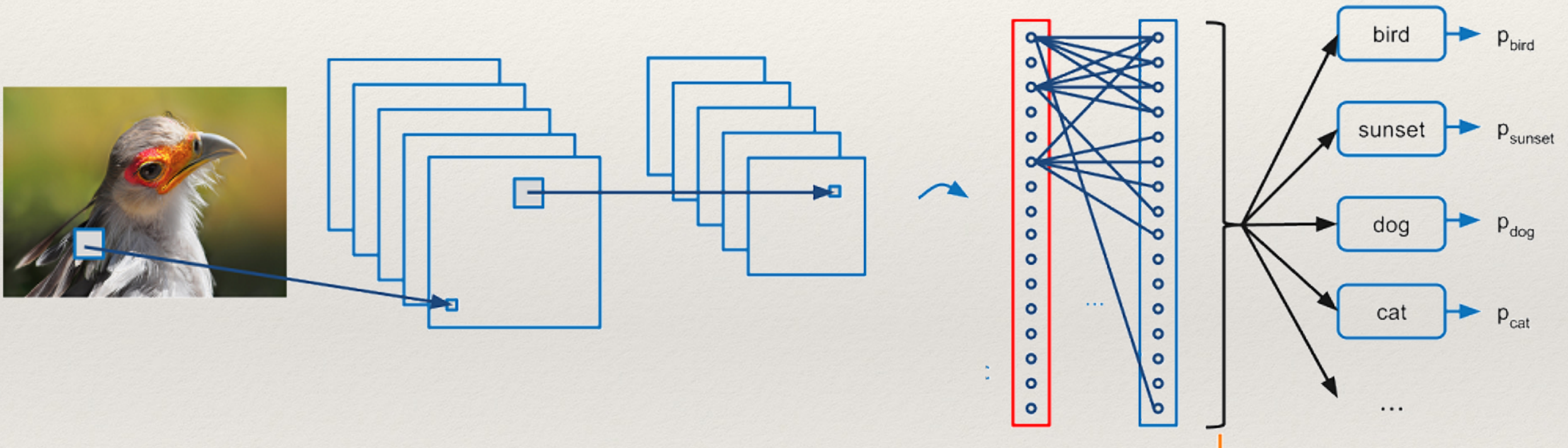
Warunkowe przepuszczanie informacji pomaga w zapamiętywaniu dalekich zależności i kontekstów



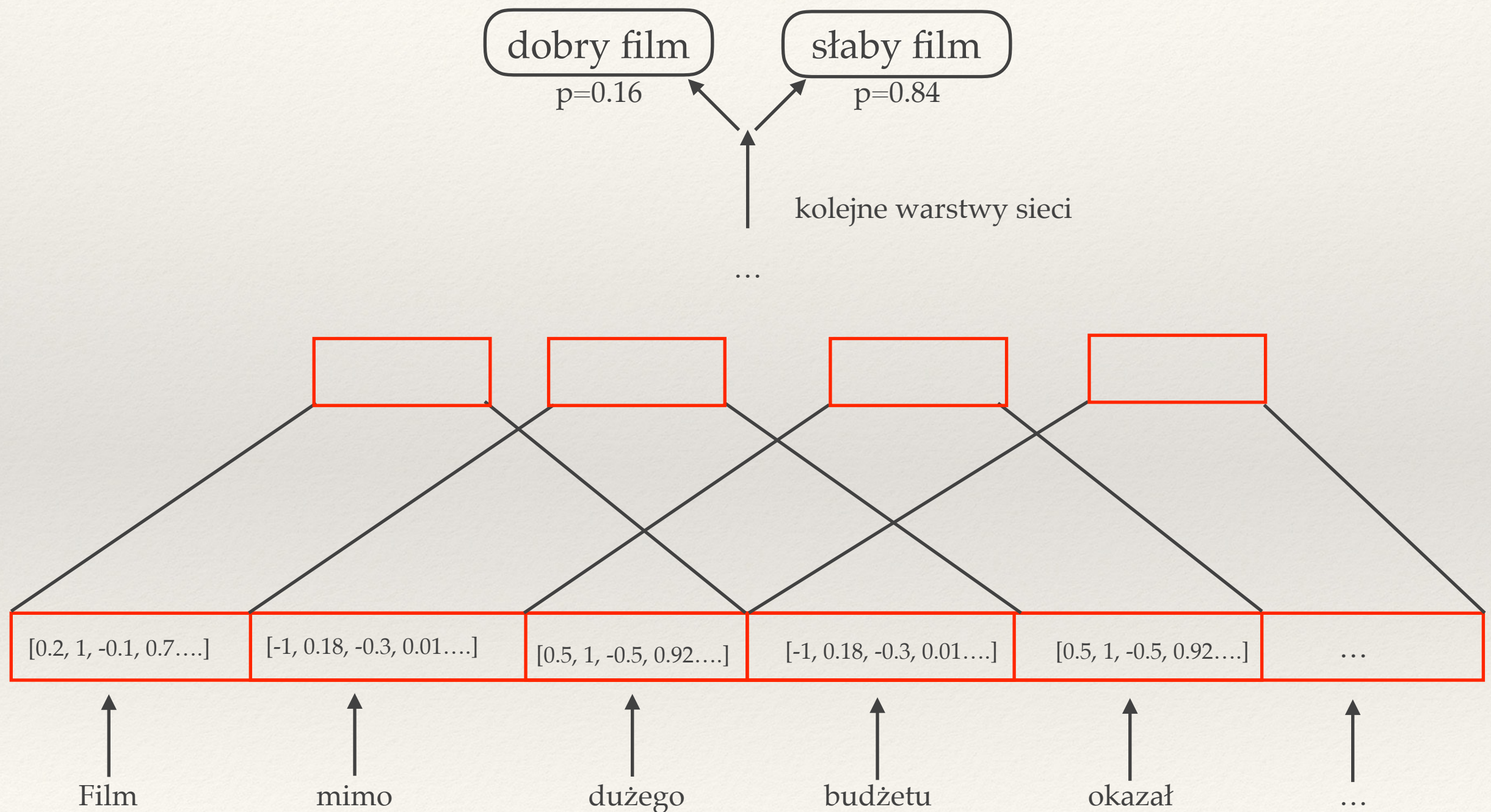
Gated Recurrent Unit

Konwolucyjne sieci neuronowe (CNN)

są bardzo skuteczne w przetwarzaniu obrazów



Jednowymiarowe CNN można użyć do analizy tekstu

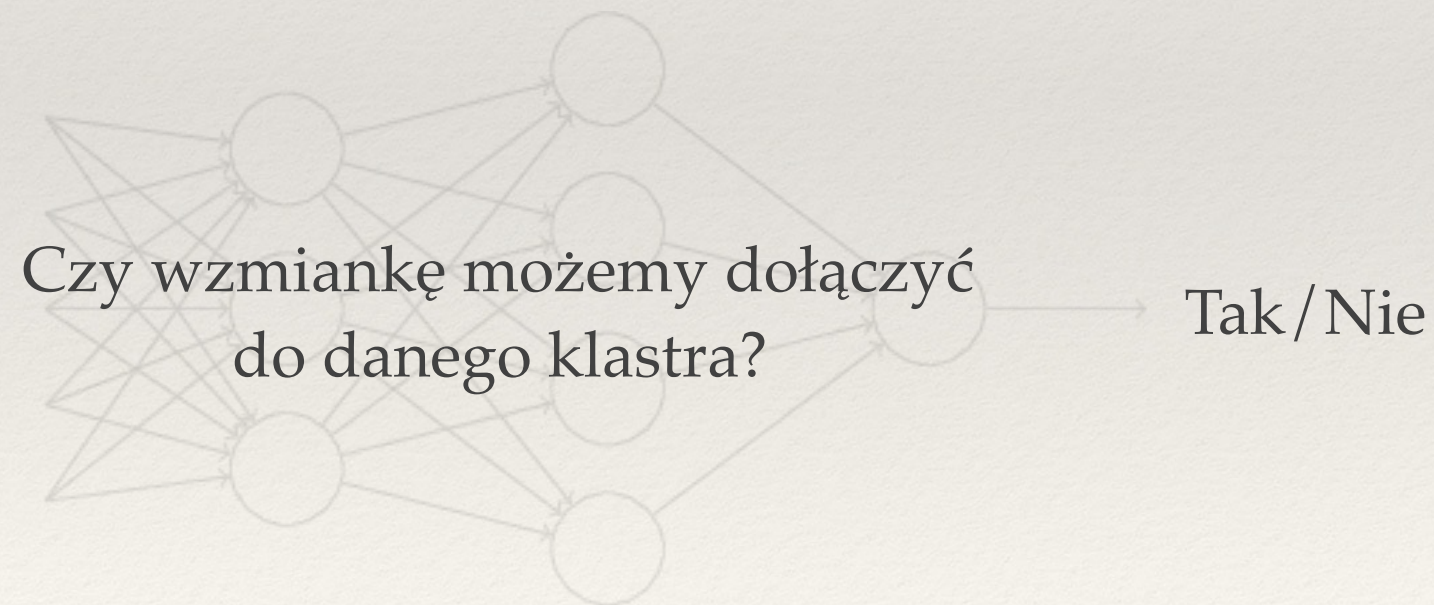
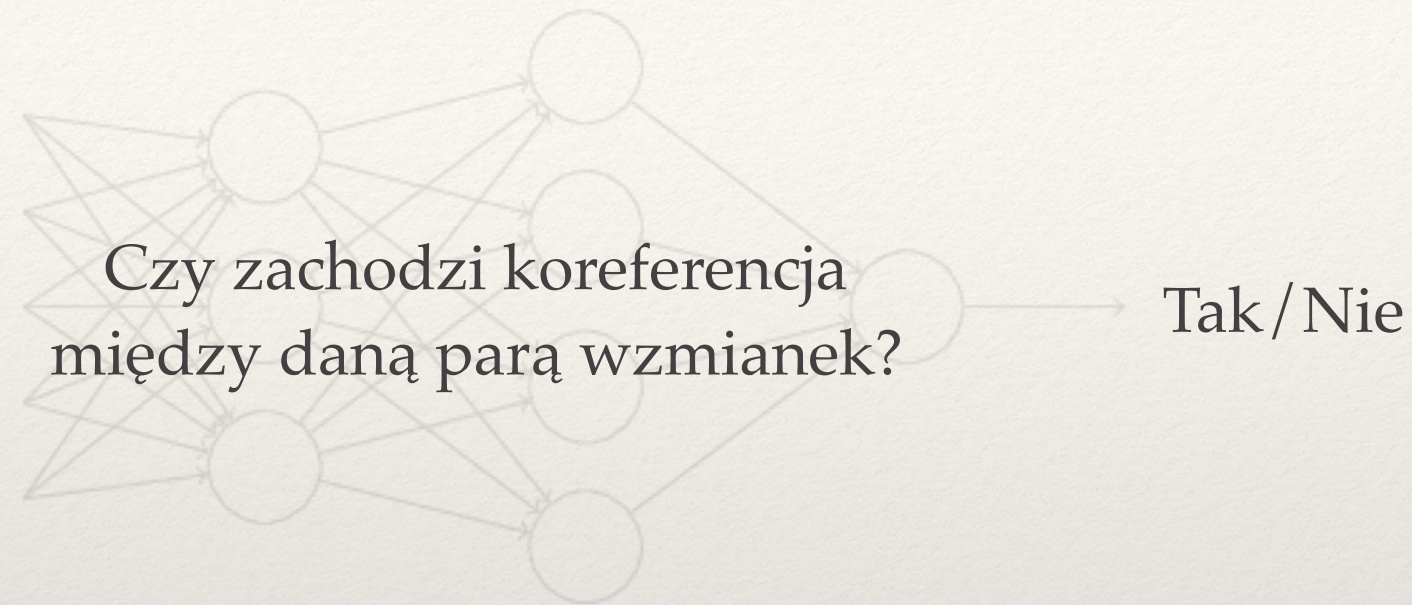


Grupowanie wystąpień koreferencyjnych nie jest łatwym zadaniem do zautomatyzowania

Był sobie raz szewczyk Dratewka. Biedak był z niego a biedak!

Aby liche odzienie na grzbiecie, aby podarte skórzniaki na nogach, aby torba, a w niej kromka chleba - cały to jego majątek.

Na bazie sieci neuronowych można zbudować system do wykrywania i grupowania koreferencji



Wykrywanie koreferencji z pomocą sieci neuronowych nie potrzebuje wielu ręcznie skrojonych cech

- ❖ Wektory (word embeddings) dla
 - ❖ głowy wzmianki, pierwszego i ostatniego słowa we wzmiance, dwóch słów przed i po wzmiance
 - ❖ uśrednione wektory dla wszystkich słów we wzmiance, wszystkich słów w zdaniu wzmianki, pięciu słów przed i po wzmiance
- ❖ Cechy relacji
 - ❖ odległość między wzmiankami
 - ❖ czy wzmianki są identyczne (przed i po lematyzacji)
 - ❖ czy wzmianki częściowo się pokrywają (overlapping)

Do uczenia sieci potrzeba dużego zbioru treningowego

- ❖ rozwiązanie ze Stanford NLP Group: około 2,5 miliona par wzmianek
- ❖ dla Polskiego Korpusu Koreferencyjnego (<http://zil.ipipan.waw.pl/PolishCoreferenceCorpus>) szacowana liczba koreferentnych wzmianek: 500 tysięcy (w zbiorze treningowym potrzebujemy również pary bez koreferencji)

Podejście do problemu koreferencji z pomocą sieci neuronowych to bogaty problem badawczy

- ❖ czy rozmiar wektorów ma znaczenie (50, 200, 300)?
- ❖ czy sposób trenowania miał znaczenie (Word2Vec, fastText)?
- ❖ czy inaczej budować reprezentację całych zdań (RNN? CNN?)
- ❖ czy dodatkowe cechy istotnie poprawią system?
- ❖ kolejność porównywania wzmianek ma znaczenie