Rozpoznawanie relacji sematycznych w tekscie, porzez klasyfikację kontekstu relacji przy pomocy klasyfikatora **CRF**

Jakub A. Gramsz
Michał Krautforst

24 stycznia 2014

Spis treści

1

Wstęp

1.1 Klasyfikator CRF

1.1.1 Motywacja

Klasyfikator utworzony pod kątem segmentacji i znakowania danych sekwencyjnych.
Często wykorzystywane w przetwarzaniu języka naturalnego.
Stosunkowo młoda metoda klasyfikacji, wykazująca potencjał do badań.
Wysoka jakość klasyfikacji otrzymana przy oznaczaniu nazw własnych w tekstach.

1.1.2 Charakterystyka klasyfikatora

Klasyfikator CRF został zaporoponowany aby przetwaorzać sekwęcje w obu kierunkach [3].

1.2 Rozpoznanie literatury

2

Praca badawcza

2.1 Implementacja

Poniżej przedstawiamy zadania.

1. Wydobycie par hiponim-hiperonim o zadanej odległości między sobą wprost z bazy Słowosieci.
2. Budowa grafu relacji hiperonimi.
3. Oznaczenie otrzymanych par w zadanym korupusie.
4. Podział korpusu na zbiór uczący oraz testujący.
5. Wydobycie oznaczonych kontekstow.
6. Wypisywanie przykładów false positive skrypt wypisujacy przyklady false positive.
7. Dobór cech dla klasyfikatora **CRF**.
8. Wyznacznanie odległości pomiędzy parami oznaczonymi poprzez klasyfikator (pośrednio oznaczonymi poprzez kontekst).

2.2 Badania

2.2.1 Zbiór uczący

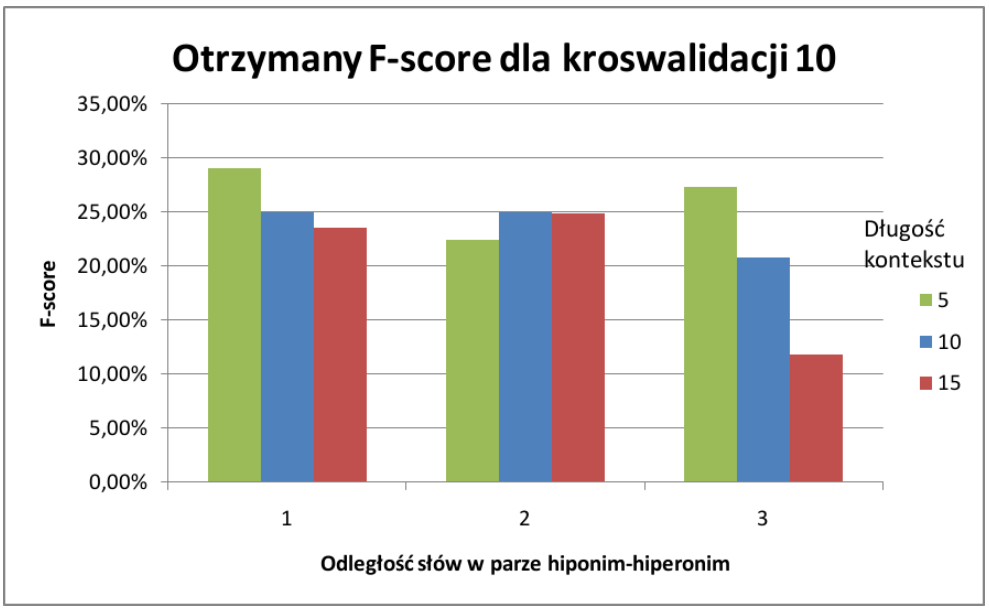
Zbiór uczący stanowią zdania pochodzące z korpusów języka polskiego z oznaczonymi kontekstami par słów znajdujących się w relacji hiperonimi znajdujących się w różnych odległościach od siebie (1 do 3) w słowosieci. Każdy token opisany wektorem cech.

- walidację krzyżową dla 2, 3 i 10 foldów.
- dla długości kontekstów 5, 10, 15
- dla par hiponim-hiperonim o odległościach 1, 2, 3.

3

Wyniki i wnioski

3.1 Wyniki



Wykres 3.1:

3.2 Wnioski

Skorzystanie z par hiponim-hiperonim o większej odległości powoduje spadek precyzji bez zmian kompletności

Długość kontekstu wpływa na liczbę wygenerowanych przypadków pozytywnych (najlepszy wynik dla 10, dłuższe konteksty nie polepszają już wyników)

Manipulacja cechami niezauważalnie wpływa na wyniki (cech powinno być możliwie dużo)

Zmniejszenie liczby negatywnych przypadków pogarsza jakość klasyfikacji. Należałoby zrównoważyć proporcje liczby przypadków pozytywnych i negatywnych, jednak trudne jest określenie kryterium.

Klasyfikacja cechuje się wysoką precyzją, mimo zróżnicowanych przypadków pozytywnych oraz rzadkiej powtarzalności kontekstów

Zdarza się, że wyodrębnione do klasyfikacji konteksty niejednoznacznie wskazują na występowanie relacji hiperonimi, przez co pomimo poprawnego działania klasyfikatora, generuje on niepoprawną parę hiponim-hiperonim

Otrzymywane wyniki walidacji przypadki False Positive par hiponim-hiperoniom zazwyczaj są pojęciami w jakiś sposób powiązanymi w odczuciu użytkownika języka naturalnego, jednak do ich oceny niezbędna jest szeroka wiedza lingwistyczna

Przykładowe FP:
- muzyk - lider
- opinia - historia
- zdarzenie - czas
- dziedzictwo - konwencja

Zaproponowana metoda odnajdowania relacji sematycznych w tekście poprzez oznaczanie ich kontekstów, nie przynosi zadowalających rezultatów, niemniej jednak może stanowić dobry wstęp na podstawie, którego innymi metodami można by proces ten kontynuować. Przykładowo oznaczone konteksty można by grupować na podstawie ich podobieństwa a następnie wyciągać z nich bardziej ogólne wzorce.

Bibliografia

[1] M. Banko, O. Etzioni, and T. Center. The tradeoffs between open and traditional relation extraction. In *ACL*, volume 8, pages 28–36, 2008.

[2] A. Culotta, A. McCallum, and J. Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303. Association for Computational Linguistics, 2006.

[3] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[4] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *IJCAI*, volume 7, pages 2862–2867, 2007.

[5] H. M. Wallach. Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22, 2004.