Aleksander Pohl

Automatyczna analiza składniowa haseł w słowniku dwujęzycznym

Niniejszy artykuł dotyczy problemu pozyskiwania wysokiej jakości informacji leksykalnej z elektronicznych słowników dwujęzycznych. We wstępie jest przedstawiona motywacja do badań nad słownikami dwujęzycznymi oraz ich krótka charakterystyka. W części drugiej został omówiony jeden z największych słowników bilingwalnych dostępnych na polskim rynku: Wielki Multimedialny Słownik Angielsko-Polski Polsko-Angielski Oxford/PWN. W części trzeciej została pokrótce omówiona technika parsowania tego rodzaju słowników oraz narzędzie, które ją wykorzystuje. W części czwartej zaś zgromadzono zalecenia dla lingwistów trudniących się tworzeniem słowników bilingwalnych, dzięki którym proces pozyskiwania z nich informacji leksykalnej może być znacznie uproszczony.

1. Wstęp

Wraz z popularyzacją technologii informacyjnych na rynku zaczęły być dostępne różnego rodzaju słowniki elektroniczne, zarówno monolingwalne, jak i bilingwalne. W szczególności, ze względu na dużą popularność języka angielskiego, dostępnych jest wiele elektronicznych słowników angielsko-polskich i polsko-angielskich, spośród których na największą uwagę zasługują Wielki Multimedialny Słownik Angielsko-Polski Polsko-Angielski Oxford/PWN (w skrócie Słownik Oxford/PWN) oraz Nowy Słownik Fundacji Kościuszkowskiej. Wersje elektroniczne tych słowników mają oczywiście znaczną przewagę nad ich wersjami papierowymi – dla zwykłego użytkownika najważniejszą zaletą jest możliwość szybkiego wyszukiwania haseł.

Słowniki elektroniczne mają znacznie większą wartość niż ich papierowe odpowiedniki, także dla naukowców zajmujących się tak zwanym przetwarzaniem języka naturalnego (*Natural Language Processing* – w skrócie NLP). Trudno sobie wyobrazić na przykład system dokonujący automatycznego tłumaczenia tekstów z języka angielskiego na polski, który nie posiłkowałby się elektronicznym słownikiem bilingwalnym[1]. Podobnie opracowanie np. polskiego WordNetu na podstawie jego angielskiej wersji można znacznie przyspieszyć dzięki wykorzystaniu takiego słownika.

[1] Dotyczy to również systemów opartych na metodach korpusowych, ponieważ odpowiedni słownik jest niezbędny m.in. na etapie uczenia algorytmu, do rozpoznania odpowiadających sobie fragmentów tekstu.

Pomimo wielkiej przydatności elektronicznych słowników bilingwalnych w dziedzinie NLP bezpośrednie wykorzystanie publicznie dostępnych słowników napotyka liczne trudności. Zasadniczy problem polega na tym, że pierwowzorami słowników elektronicznych są zazwyczaj słowniki papierowe i podstawowym kryterium, którym kierują się redaktorzy dokonujący zamiany jednej wersji na drugą, jest ich *wizualna* zgodność [por. Żmigrodzki 2008, s. 99]. Największe słowniki, poza prostym mapowaniem pomiędzy słowami obu języków, zawierają szereg informacji dodatkowych, obejmujących m.in. fonetykę, syntaktykę oraz semantykę. Do oznaczenia tych dodatkowych informacji są stosowane różne konwencje typograficzne, które, choć zrozumiałe dla użytkownika, stanowią poważny problem dla algorytmów, których celem jest wydobycie tych informacji. Pomimo że często wykorzystuje się języki znaczników, takie jak SGML czy XML, zamiast wykorzystać ich potencjał do semantycznego oznaczania poszczególnych składników hasła, zazwyczaj ogranicza się ich zastosowanie do przeniesienia odpowiednich cech wizualnych.

Co więcej, zdobycze współczesnej techniki informatycznej są rzadko wykorzystywane do weryfikacji spójności danych występujących w słowniku. W trakcie dogłębnej analizy zawartości słowników wychodzi na jaw wiele niespójności, zarówno w ich treści, jak i strukturze haseł, a nawet indeksu. Najbardziej oczywisty postulat, który można wysunąć wobec dwóch słowników bilingwalnych tego samego producenta, czyli symetryczność, nie jest spełniony wobec treści tych słowników (np. w wersji angielsko-polskiej występuje para słów *aboveboard* – *uczciwy*, a w wersji polsko-angielskiej, w haśle *uczciwy* nie występuje słowo *aboveboard*), struktury haseł, a nawet struktury indeksu (np. w wersji angielsko-polskiej w indeksie pojawiają się wyrażenia złożone, a w wersji polsko-angielskiej nie).

2. Wielki Multimedialny Słownik Angielsko-Polski Polsko-Angielski Oxford/PWN

Problemy, o których mowa w poprzedniej części, będą zilustrowane na przykładzie Słownika Oxford/PWN. Część z nich została omówiona w pracy Jassema [2003], jednakże autor tego opracowania dysponował nieco odmienną wersją słownika oraz koncentrował się na niespójnościach występujących na wyższym poziomie abstrakcji. Tutaj natomiast chcemy przyjrzeć się pewnym problemom bardziej szczegółowo.

Słownik Oxford/PWN w wersji z roku 2004 zawiera w części angielsko-polskiej ok. 63 000 haseł, a w części polsko-angielskiej ok. 52 000 haseł. Obie części zawierają słownictwo ogólne oraz specjalistyczne, hasła zaś są opisane przy użyciu języka SGML, ale ich struktura nie została publicznie udokumentowana.

mysz *f*

1. Zool. (house) mouse; mysz domowa/polna/leśna house/field/wood mouse; pisk myszy mouse's peep; pułapka na myszy a mousetrap
2. Komput. mouse; kliknąć myszą to click the mouse; podkładka do myszy a mouse pad

balk /bɔlk/

I *n*

1. Agric miedza *f*
2. Constr belka *f*, krawędziak *m*
3. US (in baseball) spalony *m* (wykonany przez miotacza)

II *vt* udaremni|ć, -ać, z|niweczyć [intention, plan, scheme]; ...

Rycina 1. Przykładowe hasła występujące w słowniku polsko-angielskim i angielsko-polskim

Na rycinie 1 są przedstawione dwa przykładowe hasła – po jednym z części angielsko-polskiej oraz polsko-angielskiej, a na rycinie 2 – fragment wewnętrznej struktury hasła.

```
<BIG><B><PL>mysz</PL></B></BIG> <I>f</I>
<P> <B>1. <HANGINGPAR></B>
<TEXTSECTION ID="1"><SMALL>Zool.</SMALL>
<TEXTSECTION ID="0">
<GB>(house) mouse</GB>;
<B><PL>mysz domowa/polna/leśna</PL></B>
<GB>house/field/wood mouse</GB>;
<PL><B>pisk myszy</B></PL>
<GB>mouse&rsquo;s peep</GB>;
<PL><B>pułapka na myszy</B></PL>
<GB>a mousetrap</GB></P>
<P> <B>2. <HANGINGPAR></B>
<TEXTSECTION ID="1"><SMALL>Komput.</SMALL>...
```

Rycina 2. Szczegółowa struktura hasła

Jak widać na pierwszej rycinie, poza prostym odwzorowaniem słów jednego języka na drugi, w hasłach występuje wiele dodatkowych informacji leksykalnych, m.in. transkrypcja fonetyczna, kategoryzacja gramatyczna, kwalifikatory dziedzinowe itp.

Wszystkie te informacje są niezwykle przydatne z punktu widzenia algorytmów przetwarzania języka naturalnego bazujących na słowniku tego rodzaju, dlatego wysoce pożądane jest, aby w procesie jego przekształcania do formy akceptowalnej w tych algorytmach nie zostały one pominięte ani przekłamane. Rycina 2. ujawnia

jednak zasadniczy problem stojący na przeszkodzie w osiągnięciu tego celu, w którym jest wykorzystywanie jedynie tagów służących do wizualnego opisu poszczególnych elementów wchodzących w skład hasła[2]. Dlatego też odróżnienie np. kwalifikatora dziedzinowego (np. *Agric* w haśle *balk*) od kwalifikatora syntagmatycznego[3] (np. *in baseball* w haśle *balk*) nie mogło ograniczyć się do najbliższych tagów okalających, które w obu wypadkach miały wartość <small>.

Poza tym zasadniczym problemem słownik ma wiele innych własności utrudniających jego przekształcenie do formatu akceptowalnego dla algorytmów przetwarzania języka naturalnego. Najważniejsze z nich są wymienione niżej:

1. Odmienna struktura indeksów haseł języka polskiego i angielskiego – w indeksie angielskim pojawiały się wyrażenia złożone (por. *white fish*, *white fox*), natomiast w polskim nie.
2. Odmienna struktura haseł języka polskiego i angielskiego, np.:
   a) w polskich hasłach pojawia się sekcja z wyrażeniami złożonymi, której brak w hasłach angielskich,
   b) odmiennie oznaczone kwalifikatory paradygmatyczne[4],
   c) w polskim tłumaczeniu angielskiego słowa pojawiają się kategorie gramatyczne.
3. Wieloznaczność elementów oddzielających składniki hasła, np.:
   a) przecinek – służy do separacji kwalifikatorów, tłumaczeń prostych i wariantów przykładów użycia, ale w tych ostatnich może występować jako zwykły znak przestankowy (por. *catch*, *całować*),
   b) średnik – oddziela grupy bliskoznacznych tłumaczeń prostych, przykłady użycia słów oraz alternatywne tłumaczenia przykładów (por. *label*, *cal 1*),
   c) nawiasy okrągłe – oznaczenie wyróżnionej formy gramatycznej, kwalifikatora paradygmatycznego i syntagmatycznego, alternatywnej pisowni oraz segmentu opcjonalnego (por. *label*).
4. Brak stałego schematu hasła, inna struktura dla hasła:
   a) zawierającego tłumaczenia dla wielu części mowy (por. *balk*),
   b) zawierającego tłumaczenia dla homonimów (por. *zamek*),
   c) zawierającego wyłącznie jedno tłumaczenie (por. *awanturniczo*).
5. Stosowanie wielu leksemów w jednym haśle (por. *Abisyńczyk/Abisynka*).

Zadanie wydobycia informacji leksykalnej ze Słownika Oxford/PWN sprowadzało się do rozwiązania trzech podproblemów:

1. opracowania spójnego indeksu, za pomocą którego można by odnajdywać tłumaczenia wybranego słowa polskiego/angielskiego,
2. opracowania uogólnionej struktury hasła słownikowego, która uwzględniałaby wszystkie fenomeny lingwistyczne rejestrowane przez słownik,

[2] W słowniku występują dwa tagi semantyczne <PL> oraz <GB>, służące do oznaczenia fragmentów odpowiednio w języku polski i angielskim. Jednak wszystkie pozostałe przydatne informacje leksykalne były oznaczane wyłącznie za pomocą tagów wizualnych.

[3] Przez kwalifikatory syntagmatyczne są rozumiane słowa służące do odróżnienia homonimów za pomocą relacji syntagmatycznych, np. *abandonment (of person, place) – opuszczenie.*

[4] Przez kwalifikatory paradygmatyczne są rozumiane słowa służące do odróżnienia homonimów za pomocą relacji hiperonimii bądź synonimii, np. *zamek (budowla) – castle.*

3. opracowania szybkiego parsera, który pozwoliłby na przekształcenie haseł słownikowych na strukturę wymienioną w punkcie 2.

Najtrudniejszym problemem jest oczywiście ten wyszczególniony w punkcie 3 – ze względu na wieloznaczność zarówno tagów, jak i innych elementów (np. przecinków) służących do separowania poszczególnych składników hasła. Nie mniej istotnym problemem okazało się również opracowanie mechanizmu weryfikacji działania parsera. Z jednej strony, oczywiste jest, że ręczne zweryfikowanie wyników parsowania ponad 100 000 haseł nie wchodziło w rachubę. Z drugiej strony, nie można było poprzestać na weryfikacji jedynie pewnej próbki haseł, gdyż istniało wysokie ryzyko pominięcia wielu haseł o specyficznych własnościach.

3. Algorytm parsowania i weryfikacja jego poprawności

Pomimo wielu problemów zasygnalizowanych w poprzedniej części artykułu, na początku wydawało się, że problem parsowania haseł w Słowniku Oxford/PWN nie jest aż tak trudny do rozwiązania. Pierwsze przybliżone rozwiązanie wykorzystywało deterministyczny automat skończony bez stosu [Hopcroft, Motwani i Ullman 2005], którego tablica przejść została zaprojektowana ręcznie, na podstawie analizy kilku haseł występujących w słowniku. Na wstępie, za pomocą prostego skanera zamieniano tagi oraz inne elementy strukturalne (przecinki, średniki, nawiasy itp.) na symbole parsera (np. „<BIG>" → BIG_O, „</BIG>" → BIG_C, „;" → SEMICOLON itp.), zaś pozostałe składniki strumienia wejściowego zamieniane były na symbol TEXT. Tak przekształcony strumień był podawany na wejście parsera, gdzie na podstawie kombinacji aktualnego stanu oraz symbolu wejściowego było dokonywane przejście do nowego stanu, które mogło również wyzwalać akcję wypełniania struktury parsowanego hasła (np. określenia kategorii gramatycznej na podstawie zawartości symbolu TEXT).

Początkowo tablica parsera zawierała około 20 stanów, dlatego jej ręczna modyfikacja nie nastręczała większych problemów. Jednak szybko okazało się, iż pomimo tego, że hasła w słowniu wyglądają podobnie, zawierają wiele subtelnych różnic, które istotnie wpływają na poprawność procesu parsowania, a naprawianie wykrytych błędów wyłącznie poprzez śledzenie zmian stanów parsera jest zadaniem karkołomnym.

W pierwszym etapie system został rozbudowany o moduł pozwalający na wizualne śledzenie zmian stanów parsera. Dzięki możliwości grupowania stanów w bloki funkcjonalne (np. blok odpowiedzialny za parsowanie transkrypcji fonetycznej), naprawianie tablicy parsingu stało się dużo prostsze. Następne usprawnienie polegało na możliwości bezpośredniej modyfikacji tablicy parsingu w module wizualizacyjnym. W szczególności – możliwe było dodanie nowego przejścia między stanami po zapoznaniu się z wyjściem skanera, ale przed aktywowaniem odpowiedniego przejścia. Ostatnie usprawnienie polegało na możliwości zakładania pułapek[5] na poszczególnych stanach. Dzięki temu możliwe było całkowicie interaktywne tworzenie tablicy parsingu.

[5] Jeśli dany stan miał założoną pułapkę, proces parsowania zatrzymywał się w momencie osiągnięcia tego stanu.

Problem weryfikacji poprawności parsingu został rozwiązany dzięki zastosowaniu trzech heurystyk walidacji haseł. Pierwsza z nich polegała na testowaniu miękkiej zgodności haseł z tablicą parsingu – jeśli w trakcie parsowania wystąpienie więcej niż czterech kolejnych symboli nie spowodowało przejścia do nowego stanu, sygnalizowany był błąd. Dwie pozostałe heurystyki polegały na obserwowaniu wyników parsingu. W pierwszej z nich błąd był sygnalizowany, jeśli wynikowe hasło nie miało żadnego prostego (jednowyrazowego) tłumaczenia. W drugiej natomiast błąd był sygnalizowany, jeżeli wynikowe hasło było pozbawione kategorii gramatycznej. Dzięki zastosowaniu heurystyk udało się rozpoznać wiele wariantów struktury hasła, które nie zostały zaobserwowane przy pobieżnym przeglądaniu słownika. Wyniki zastosowania heurystyk są zgromadzone w tabeli 1. i odpowiadają one liczbie haseł, które zostały wskazane jako wadliwe według danej heurystyki (np. w słowniku angielsko-polskim 0,45% wynikowych haseł było pozbawionych kategorii gramatycznej).

Otrzymane wyniki pokazują, że możliwe było uzyskanie dobrej zgodności tablicy parsingu ze strukturą hasła. Zastanawiający może być jednak brak w ok. 10% wynikowych haseł jednowyrazowych tłumaczeń, który mógłby wskazywać, że pomimo poprawnego parsowania, część informacji zgromadzonych w słowniku jest tracona. Jednakże pobieżne zbadanie haseł oznaczonych przez tę heurystykę jako błędne wykazało, że w większości przypadków w źródłowym haśle nie występowało jednowyrazowe tłumaczenie (najczęściej tłumaczenie było wyrażeniem złożonym – por. *bajda*, ale również pojawiały się hasła, w których zamiast tłumaczenia było omówienie – por. *bakalie*, a także przetłumaczone było jedynie użycie danego słowa w określonym kontekście – por. *bambuko*).

Tabela 1. Wyniki zastosowania heurystyk – liczba haseł oznaczonych jako niepoprawne

Słownik	angielsko-polski	polsko-angielski
Parsowanie (5 kolejnych symboli)	0,31%	0,55%
Proste tłumaczenia (pojedyncze słowa)	9,47%	10,92%
Kategoria gramatyczna	0,45%	8,2%

Więcej informacji na temat programu oraz jego kod źródłowy zostanie udostępnione na stronie internetowej http://apohllo.pl/projekty/viper. Przykładowy zrzut ekranu zawierający fragment tablicy parsingu słownika angielsko-polskiego jest przedstawiony na rycinie 3.

4. Zalecenia dla twórców słowników

Problem wydobywania informacji leksykalnej ze słowników bilingwalnych niewątpliwie byłby zdecydowanie prostszy, gdyby słowniki tego rodzaju projektowano z uwzględnieniem algorytmów przetwarzania języka naturalnego. W trakcie pracy nad Słownikiem Oxford/PWN zgromadzonych zostało wiele uwag, na podstawie których opracowano pewne zalecenia. Ich przestrzeganie w znaczącym stopniu ułatwi wykorzystanie słowników tego rodzaju w dziedzinie NLP.

Idealny słownik bilingwalny:

1. Wykorzystuje jeden schemat opisu hasła.
2. Zawiera wektory odmiany słów (w szczególności dla języka polskiego).
3. Bazuje na spójnych słownikach semantycznych (np. WordNecie) dla obu języków.
4. Jest symetryczny (zarówno w wymiarze treści, jak i struktury hasła oraz struktury indeksu).
5. Stosuje jednolity indeks:
   a) wszystkie leksemy o identycznej formie bazowej są reprezentowane jako jeden wpis w indeksie (odróżniane są jednak hasła pisane wielką i małą literą),
   b) nie zawiera odsyłaczy – jeśli ma kilka wariantów, wpisy w indeksie prowadzą bezpośrednio do tej samej definicji,
   c) każdy wariant hasła pojawia się w indeksie.
6. Stosuje jednolitą strukturę hasła:
   a) podział nadrzędny: obowiązkowe kategorie gramatyczne,
   b) podział wewnątrz kategorii gramatycznej na leksemy (tzn. odróżnienie słów o odmiennym wektorze odmiany, por. *zamku*, *zamka*),
   c) podział leksemu na grupy semantyczne, w oparciu o słownik monolingwalny, najlepiej elektroniczny słownik semantyczny,
   d) ścisły podział grupy semantycznej na tłumaczenia jednowyrazowe, wielowyrazowe oraz przykłady użycia,
   e) ścisły podział kwalifikatorów na kwalifikatory dziedziny, rejestru, stylu, paradygmatyczne, syntagmatyczne etc.
7. Stosuje jednoznaczne oznaczenia do separowania informacji dostępnych w ramach jednego hasła (najlepiej poprzez wykorzystanie języka XML i tagów semantycznych).
8. Przechowuje wyrażenia złożone wyłącznie w indeksie (por. *pułapka na myszy*, która w Słowniku Oxford/PWN pojawia się jedynie w definicja hasła *mysz*, ale brakuje jej w definicji hasła *pułapka*).

Przestrzeganie wyżej wymienionych zaleceń w konstrukcji słowników bilingwalnych z pewnością przyczyni się do ich łatwiejszej adopcji w dziedzinie NLP. Jednocześnie narzucenie rygorystycznych wymagań co do struktury i zawartości słownika może również przyczynić się do zwiększenia jego przydatności dla zwykłych użytkowników, m.in. dzięki poprawieniu dostępności informacji zgromadzonych w słowniku. Na szczęście sygnalizowane tutaj problemy zaczynają być dostrzegane przez samych leksykografów i niektóre z powyższych zaleceń są brane pod uwagę przy konstrukcji nowych słowników elektronicznych (niekoniecznie bilingwalnych), np. Wielkiego Słownika Języka Polskiego [Żmigrodzki 2008, s. 114–123].

BIBLIOGRAFIA

Hopcroft J.E., Motwani R., Ullman J.D. (2005). *Wprowadzenie do teorii automatów, języków i obliczeń*, tłum. B. Konikowska. Warszawa: Wydawnictwo Naukowe PWN.

Jassem K. (2004). *Applying Oxford-PWN English-Polish Dictionary to Machine Translation*. Proceedings of the 9th EAMT Workshop, „Broadening horizons of machine translation and its applications", s. 98–105.

Wielki Multimedialny Słownik Angielsko-Polski Polsko-Angielski Oxford/PWN (2004). Warszawa: Wydawnictwo Naukowe PWN.

Żmigrodzki P. (2008). *Słowo – słownik – rzeczywistość*. Kraków: Lexis.