Wytyczne lematyzacji fraz rzeczownikowych i przymiotnikowych

Adam Radziszewski, Marcin Oleksy, Jan Wieczorek
Instytut Informatyki, Politechnika Wrocławska

Grudzień 2012

Wprowadzenie

Lematyzacja fraz polega na przypisaniu wystąpieniom fraz w tekście ich **form podstawowych**. Forma podstawowa frazy jest poprawną frazą tego samego typu (nie chodzi o sprowadzenie wszystkich wyrazów do form podstawowych), która mogłaby wystąpić w słowniku albo na liście słów kluczowych.
Lematyzacja zawsze (są wyjątki związane z nazwami własnymi) wymaga sprowadzenia nadrzędnika frazy do mianownika, na ogół wymaga też sprowadzenia go do liczby pojedynczej, a czasem również rodzaju do męskiego.
Rozpatrujemy tutaj frazy NP (zgodnie z wytycznymi KPWr, frazą typu NP są też frazy przyimkowe) i AdjP.
Przykłady wprowadzające:

1. [chorobami cywilizacyjnymi] -> choroba cywilizacyjna
2. [Ministerstwa Edukacji Narodowej] -> Ministerstwo Edukacji Narodowej
3. [nas] -> my
4. [przez łąkę i las] -> łąka i las
5. [przywiązaną do tradycji] -> przywiązany do tradycji

Ogólne założenia

1. Rozróżniamy wielkość liter. Lematyzacja wiąże się też z normalizacją wielkości liter — jeśli pierwsze słowo frazy jest wyrazem pospolitym, forma podstawowa powinna zaczynać się małą literą, niezależnie od oryginalnej pisowni (fraza napotkana mogła rozpoczynać zdanie). Jeśli obie pisownie są sensowne, preferujemy pisownię małą literą.

2. Znaczna część fraz jest już w formie podstawowej i nie wymaga zmian. Często też jedyną wymaganą zmianą jest sprowadzenie do małej litery.
3. Skrótów nie rozwijamy. Przykładowo, nie sprowadzamy [prof.] do [profesor]. Skróty mogą jednak ulec pewnym zmianom, w szczególności może zajść konieczność obcięcia końcówki fleksyjnej z akronimu (np. [PWN-u] -> [PWN]).

Lematyzacja fraz rzeczownikowych

Lematyzacja fraz rzeczownikowych wymaga sprawdzenia 3 kategorii gramatycznych **nadrzędnika**: przypadka, liczby i rodzaju:

1. Przypadek zawsze sprowadzamy do **mianownika**.
2. Liczbę na ogół sprowadzamy do **liczby pojedynczej**. Liczba pozostaje mnoga w przypadku nazw własnych i tytułów (Milionerzy), plurale tantum (nożyczki) i wtedy, gdy zmiana liczby zmienia znaczenie (warunki uzyskania przychodów, szanse), bądź nakazuje tego uzus (fraza sprowadzona do liczby pojedynczej brzmi źle). Można wyobrazić sobie, że formy podstawowe fraz są słowami kluczowymi podsumowującymi tekst i niektóre z nich będą naturalniej brzmiały w liczbie mnogiej. Ścisłego kryterium nie sposób sformułować, zobaczymy rozbieżności anotacji.
3. Rodzaj rzeczownika zawsze zachowujemy.

Jeśli fraza zaczyna się **przyimkiem**, należy przyimek odciąć przed lematyzacją. Przyimek też może być wtórny, np. [ze względu na wysoką stopę procentową] -> wysoka stopa procentowa. Wyjątkiem są nazwy własne, tytuły itp., np. [W pustyni i w puszczy] -> W pustyni i w puszczy.

Lematyzacja fraz przymiotnikowych

W przypadku fraz przymiotnikowych rozpatrujemy 4 kategorie gramatyczne **nadrzędnika**: przypadek, liczbę, rodzaj i **stopień**:

1. Przypadek zawsze sprowadzamy do mianownika.
2. Liczbę sprowadzamy do liczby pojedynczej (dotyczy nadrzędnika!). Wyjątkiem są przymiotniki użyte nominalnie, które chcemy potraktować na równi z rzeczownikami. Wyjątkiem są też nazwy własne i tytuły (np. Szybcy i wściekli).
3. Rodzaj sprowadzamy do **męskiego**. Wyjątkiem są nazwy własne i tytuły (np. Rozważna i romantyczna).
4. **Stopień zachowujemy**.

Zaimki osobowe

Zachowujemy liczbę i rodzaj zaimka. Sprowadzamy jedynie przypadek do mianownika. Przykładowo: wami -> wy, ty -> ty, niego -> on, niej -> ona, nimi -> **one lub oni**.
Jeśli na podstawie dostępnego kontekstu nie da się ustalić rodzaju, preferujemy męskoosobowy.

Zaimki dzierżawcze i wskazujące

Jeśli zaimek nie jest nadrzędnikiem frazy, należy jego formę dopasować do wymagań składniowych narzuconych przez nadrzędnik, np. jego owcami -> jego owca, naszymi owcami -> nasza owca, tymi owcami -> ta owca.
Zaimki wskazujące zachowują się jak przymiotniki i tak też je traktujemy — sprowadzamy rodzaj do męskiego, przypadek do mianownika, a liczbę do pojedynczej. Sprowadzamy więc je zawsze do „ten”, „tamten”, „taki” itp. (w przypadku wskazujących będących nadrzędnikiem frazy).
Zaimki względne o formach przymiotnikowych (jaki, który) traktujemy jak wskazujące.
Zaimki dzierżawcze można podzielić na dwie grupy: odmieniające się jak przymiotniki (mój, twój, nasz, wasz) i odmieniające się jak rzeczowniki (jego, jej, ich).
Formy przymiotnikowe zaimków dzierżawczych traktujemy tak samo jak zaimki wskazujące — sprowadzamy je zawsze do mianownika liczby poj. rodzaju męskiego, np. „moimi” -> „mój”. Zasady postępowania są więc identyczne jak w przypadku fraz przymiotnikowych.
Zaimki odmieniające się jak rzeczowniki (jego, jej, ich) mają rodzaj i liczbę przypisany na stałe i rodzaj ten należy zostawić, np. [ich] -> oni lub one, jej -> ona, jego -> on.
Jeśli na podstawie dostępnego kontekstu nie da się ustalić rodzaju, preferujemy męskoosobowy.
Zmiana: zaimki wskazujące i dzierżawcze o odmianie przymiotnikowej traktujemy tak samo, czyli po przymiotnikowemu (liczba pojedyncza, rodzaj męski). Wcześniej było niespójnie.