Spejd Web Service

Szymon Acedański
Uniwersytet Warszawski
accek@mimuw.edu.pl

Abstrakt

W artykule opisano Web Service służący do powierzchniowego parsowania tekstów przy pomocy Spejda — narzędzia do tego celu opracowanego w IPI PAN polskiego opracowanego w IPI PAN. Przeniesienie Spejda w świat internetu ma służyć przede wszystkim ułatwieniu korzystania z niego przez jak najwięcej osób. Dlatego w tej pracy skupiono się na opisie usługi od strony użytkownika, przedstawieniu możliwości i sposobu użycia.

Spejd Web Service od wersji 3 wykorzystuje do oznaczania wprowadzonego czystego tekstu tagger TaKIPI (Piasecki, 2007) przy pomocy dostarczonego Web Service'u (Piasecki, 2009).

1 Wstęp

Spejd (Przepiórkowski and Buczyński, 2007) to narzędzie do powierzchniowego przetwarzania tekstów oraz do dezambiguacji morfosyntaktycznej. Został opracowany w IPI PAN w 2007 roku w ramach prac nad projektem Narodowego Korpusu Języka Polskiego.

Stworzenie Web Servicu do tego programu ma na celu ułatwienie korzystania z niego jak największej rzeszy użytkowników. Ma również zachęcić do tworzenia łatwo dostępnych, bo opartych na technologii internetu, serwisów pokrewnych, udostępniając prosty i przejrzysty interfejs programistyczny. Wykorzystanie web servicu eliminuje również nakład pracy związany z koniecznością dołączania i utrzymywania Spejda przez autorów (względnie użytkowników) programów, które z niego korzystają.

Niejako przy okazji powstały również dwie dodatkowe rzeczy:

- prosta strona internetowa pod nazwą *iSpejd*, przez którą można interaktywnie korzystać z systemu — wystarczy wprowadzić zestaw reguł oraz tekst (można też wybrać reguły bądź teksty z listy predefiniowanych), żeby w przeglądarce zobaczyć efekt przetwarzania,
- skrypt i usługa służące do konwersji dokumentów z formatu IPI PAN do HTMLa.

W kolejnych częściach pracy opisuję pokrótce sposób korzystania z usługi Spejd Web Service oraz strony iSpejd.

2 Dostęp do usługi Spejd

Do komunikacji z usługą sieciową Spejda używany jest protokół XML-RPC (Winer, 1999). Usługa działa pod adresem `http://spejdws.appspot.com/xmlrpc`. Dokładny opis składni zapytań oraz sposobu ich wysyłania znajduje się w specyfikacji wspomnianego protokołu.

Usługa Spejd Web Service udostępnia następujące operacje:

- `string getVersion()`

  Zwraca napis zawierający numer wersji usługi.

- `list<string> listPredefinedRuleSets()`

  Zwraca listę nazw predefiniowanych zestawów reguł zapisanych w systemie.

  Można ich użyć w operacji `parse` przekazując jako parametr *rulesMode* napis `PREDEFINED`, a nazwę reguły jako parametr *rules*.

- `string getPredefinedRuleSet(string name)`

  Zwraca reguły z predefiniowanego zestawu o wskazanej nazwie. Są zwracane jako pojedynczy napis (zazwyczaj zawierający wiele linii) w języku Spejda.

- `list<string> listPredefinedTexts()`

  Zwraca listę nazw predefiniowanych tekstów.

  Można ich użyć w operacji `parse` przekazując jako parametr *textMode* napis `PREDEFINED`, a nazwę tekstu jako parametr *text*.

- `string getPredefinedText( string name)`

  Zwraca predefiniowany tekst o danej nazwie. Jest on zwracany jako dokument XML w formacie IPI PAN.

  Modyfikacja predefiniowanych zestawów reguł oraz tekstów wymaga interwencji operatora. Jest opisana w pliku `README` w dystrybucji serwisu.

- `string parse(string rulesMode, string textMode, String rules, string text)`

  Parsuje podany tekst za pomocą podanych reguł i zwraca wynik jako dokument XML w formacie IPI PAN.

  Jeśli jako *rulesMode* przekazano napis `PREDEFINED`, wtedy w parametrze *rules* należy przekazać nazwę predefiniowanego zestawu reguł. Jeśli się chce użyć własnych reguł, należy jako parametr *rulesMode* przekazać wartość `CUSTOM`.

  Podobnie parametr *textMode* może przyjąć wartość `PREDEFINED` jeśli chce się użyć gotowego tekstu. Jeśli chce się podać własny tekst w formacie IPI PAN, należy podać wartość `XML`. Usługa akceptuje również zwykły tekst. Wtedy należy podać `PLAIN`.

  Istnieje też możliwość przepuszczenia dostarczonego czystego tekstu przez tager morfosyntaktyczny TaKIPI przed przekazaniem do Spejda. W tym przypadku jako *testMode* należy przekazać `PLAIN-TO-TAG`.

  Czysty tekst nie jest przetwarzany analizatorem morfosyntaktycznym. Jest poddawany prostej segmentacji: granice zdań wyznaczają kropki, a segmentów wyrazy. Wyrazom przypisywany jest jeden możliwy tag — IGN. Znaki przestankowe są traktowane jako osobne segmenty otagowane jako INTERP.

  Do przetwarzania tekstów wykorzystana jest domyślna konfiguracja Spejda oraz domyślna konfiguracja tagsetu, odpowiadająca tagsetowi IPI PAN (Przepiórkowski and Woliński, 2003) z dodanymi dwoma specjalnymi klasami gramatycznymi: LICZBA oraz WALUTA.

Od wersji 1.1 usługa udostępnia jeszcze dwie funkcje, dzięki którym można jej użyć do konwersji dowolnych tekstów w formacie IPI PAN do eleganckiego HTMLa.

- `string formatXmlAsHtml( string xml)`

  Konwertuje przekazany dokument z formatu IPI PAN do formatu HTML. Zwraca fragment zgodny ze standardem HTMLa. Warto zwrócić uwagę, że zwrócony dokument nie jest sam w sobie poprawnym dokumentem HTML, lecz jedynie jego fragmentem, który należy umieścić wewnątrz elementu `<body>`.

  Aby poprawnie wyświetlić zwrócony fragment, następujący plik CSS musi zostać dołączony do dokumentu: `http://spejdws.appspot.com/css/ipipanxces.css`.

  Wygodny skrypt, którego można użyć z linii komend do wyświetlania tekstów sformatowanych tym sposobem, znajduje się w dystrybucji w katalogu `extras`. Nazywa się `ipipan2html.py`.

- `string parseAndFormatAsHtml( string rulesMode, string textMode, string rules, string text)`

  Ta funkcja jest po prostu połączeniem dwóch powyższych. Parsuje podany tekst za pomocą wybranych reguł oraz zwraca wynik w formacie HTML.

Przy wywoływaniu powyższych metod przez XML-RPC, ich nazwy należy poprzedzić prefiksem „`SpejdService.`”. Listing 1 pokazuje przykładowe zapytanie. Metodą `parse` przetwarzany jest tekst „Jest 2000 rok.” przy użyciu jednej reguły, która ustawia tag wszystkich liczb na LICZBA oraz tworzy z nich jednoelementowe grupy składniowe.

Listing 1: Przykład zapytania XML-RPC

```
POST /xmlrpc HTTP/1.0
Host: spejdws.appspot.com
User-Agent: xmlrpclib.py/1.0.1
    (by www.pythonware.com)
Content-Type: text/xml
Content-Length: 433

<?xml version='1.0'?>
<methodCall>
<methodName>SpejdService.parse
  </methodName>
<params>
<param>
<value>
<string>CUSTOM</string>
</value>
</param>
<param>
<value>
<string>PLAIN</string>
</value>
</param>
<param>
<value><string>
Rule "liczby"
Match: [orth~"[0-9]*"];
Eval: set(liczba,"liczba",1);
      group("LICZBA",1,1);
</string></value>
</param>
<param>
<value><string>Jest 2000 rok.
  </string></value>
</param>
</params>
</methodCall>
```

Listing 2: Odpowiedź serwera XML-RPC

```
HTTP/1.0 200 OK
Content-Type: text/xml
Date: Mon, 07 Sep 2009
  04:16:57 GMT
Server: Google Frontend
Cache-Control: private,
  x-gzip-ok=""

<?xml version="1.0"
  encoding="UTF-8"?>
<methodResponse xmlns:ex=
  "http://ws.apache.org/xmlrpc
␣␣/namespaces/extensions">
<params><param><value>

[ entity-escaped response here ]

</value>
</param></params>
</methodResponse>
```

3 *iSpejd* — interaktywny Spejd

iSpejd jest prostym serwisem WWW, przez który każdy może spróbować swoich sił w przetwarzaniu powierzchniowym. Jest dostępny na stronie `http://spejdws.appspot.com`. Po wpisaniu tego adresu w przeglądarce pojawia się formularz, w którym można wpisać bądź wybrać z listy zarówno zestaw reguł Spejda, jak również tekst do przetwarzania. Na rysunku 1 przedstawiono wygląd strony z pokazanym wynikiem parsowania.

Strona jest dostępna w dwóch językach – polskim i angielskim.

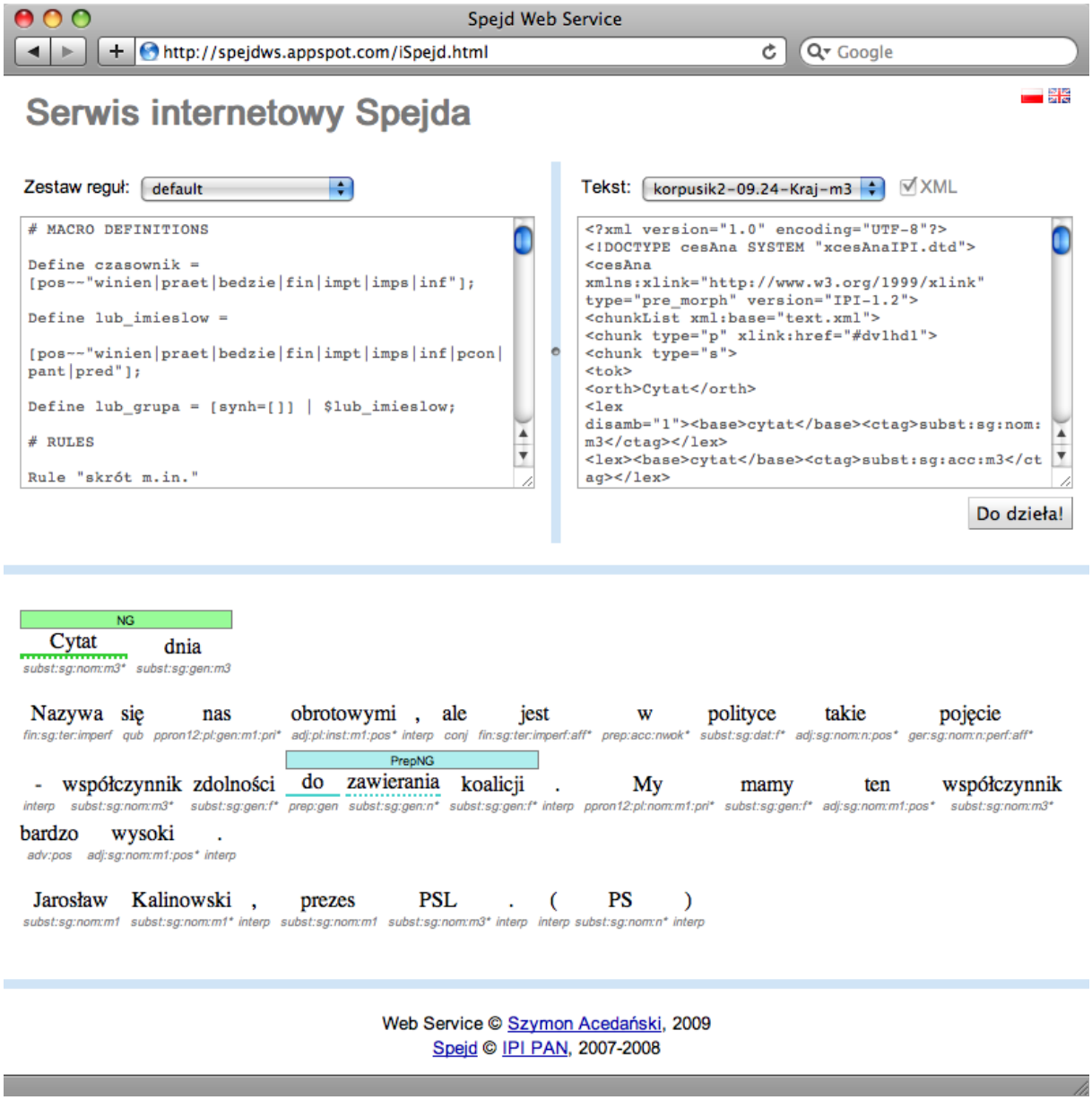
4 Informacje techniczne

Zarówno Web Service, jak i iSpejd, zostały uruchomione przy użyciu infrastruktury Google App Engine (Google Inc., 2008). Google udostępnia wszystkim swoim użytkownikom możliwość umieszczania na serwerach aplikacji internetowych. Usługa jest darmowa, o ile nie przekracza się niemałych w sumie dziennych limitów określających dopuszczalną liczbę zapytań, zużycie procesora oraz dysku itp.

Web Service został wykonany przy użyciu technologii Java Servlets (Sun Microsystems Inc., 1994). Do przetwarzania zapytań XML-RPC wykorzystana została bibliotekę *ws-xmlrpc* (The Apache Software Foundation, 2001). Strona iSpejd została wykonana przy użyciu biblioteki Google Web Toolkit (Google Inc., 2006).

5 Podsumowanie

Przedstawiony w pracy system pozwala szerokiej gamie osób zainteresowanych przetwarzaniem języka polskiego, na prosty dostęp do parsera powierzchniowego Spejd. Może być także pomocą w prowadzeniu zajęć edukacyjnych i popularyzacji dyscyplin związanych z przetwarzaniem języka naturalnego.



Rysunek 1: Strona iSpejda

Poniżej przedstawiam kilka pomysłów na dalszy rozwój serwisu.

- integracja modyfikacji Spejda wykonanych na potrzeby projektu z kodem autora,
- automatyczna ewaluacja gramatyki przy danym korpusie wzorcowym,
- parsowanie fragmentów korpusu IPI PAN wybieranych zapytaniami Morfeusza.

System można pobrać ze strony `http://www.mimuw.edu.pl/~accek/spejdws/`.

Bibliografia

Google Inc. 2006. Google Web Toolkit. `http://code.google.com/webtoolkit/`. retrieved 2009-09-06.

Google Inc. 2008. Google App Engine. `http://code.google.com/appengine/`. retrieved 2009-09-06.

Sun Microsystems Inc. 1994. Java Servlets Technology. `http://java.sun.com/products/servlet/`. retrieved 2009-09-06.

The Apache Software Foundation. 2001. Apache XML-RPC. `http://ws.apache.org/xmlrpc/`. retrieved 2009-09-06.

Maciej Piasecki. 2007. Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1–2):151–167.

Maciej Piasecki. 2009. Tagger takipi web service. Technical report, Wrocław University of Technology.

Adam Przepiórkowski and Aleksander Buczyński. 2007. ♠: Shallow parsing and disambiguation engine. In *Proceedings of the 3rd Language & Technology Conference*, Poznań.

Adam Przepiórkowski and Marcin Woliński. 2003. A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages, *EACL 2003*.

Dave Winer. 1999. XML/RPC specification. Technical report, Userland Software.