Anotacja dialogów w projekcie LUNA

J. Rabiega-Wiśniewska, M. Marciniak, A. Mykowiecka

Instytut Podstaw Informatyki PAN

Warszawa, 23 kwietnia 2007

Plan wystąpienia

- O projekcie LUNA
- Zadania zespołów francuskiego i włoskiego
- Zadania zespołu polskiego
- Anotacja morfosyntaktyczna
- Anotacja semantyczna
- Akty dialogowe
- Podsumowanie

Projekt LUNA

LUNA – spoken Language UNderstanding in multilinguAl communication systems, `http://www.ist-luna.eu/`

- Kierownictwo projektu
  - manager: prof. Renato De Mori, Université d'Avignon
  - koordynator: Silvia Mosso, Loquendo
- Ramy czasowe projektu
  - rozpoczęcie – wrzesień 2006
  - okres – trzy lata
- Partnerzy
  - Reinisch-Westfälische Technische Hochshule Aachen
  - Università degli studi di Trento
  - France Télécom
  - CSI-Piemonte
  - Polsko-Japońska Wyższa Szkoła Technik Komputerowych
  - Instytut Podstaw Informatyki PAN

Cele projektu

Założenia projektu

W ramach projektu LUNA powstanie pakiet narzędzi rozumienia mowy (Spoken Language Understanding), który może zostać wykorzystany w serwisach dialogowych kilku języków.
Dalekosiężnym celem projektu jest komunikacja człowiek-maszyna.

Główne cele projektu:

- Zebranie korpusów (naturalnych) dialogów dla języków: francuskiego, polskiego i włoskiego
- Anotacja semantyczna dialogów
- Przygotowanie materiału do trenowania:
  - automatycznej analizy i syntezy mowy
  - systemów dialogowych

Zadania poszczególnych zespołów – język francuski

- Korpus MEDIA
  - zawiera 1250 dialogów człowiek-maszyna zbieranych w środowisku Wizard-of-Oz
  - dziedzina – telefoniczna informacja o hotelach
  - trankskrypcja dialogów jest przygotowana
  - korpus jest anotowany semantycznie

Zadania poszczególnych zespołów – język francuski

- FT Stock Exchange
  - zebrano 1650 dialogów człowiek-maszyna
  - dziedzina – giełda finansowa, zarządzanie akcjami
  - korpus zostanie uzupełniony o 1000 nowych dialogów
  - transkrypcja zebranych dialogów jest gotowa
  - transkrypcja nowych dialogów i anotacja semantyczna całości – do zrobienia
- FT Customer Support service
  - zebrano 12000 dialogów człowiek-maszyna
  - dziedzina – usługi
  - korpus zostanie uzupełniony o 20000 nowych dialogów
  - transkrypcja zebranych dialogów jest gotowa
  - transkrypcja nowych dialogów i anotacja semantyczna całości – do zrobienia

Zadania poszczególnych zespołów – język włoski

CSI Piemonte

Konsorcjum założone przez Uniwersytet i Politechnikę Turynu oraz administrację Regionu Piemontu w 1977. Obecnie zrzesza poza Założycielami dwóch członków wspierających i 49 członków zwykłych.

- CSI Customer Support i Uniwersytet Trento
  - zostanie zebrany korpus 500 dialogów człowiek-człowiek
  - dziedzina – help desk, usługi
  - transkrypcja i anotacja semantyczna dialogów – do zrobienia
  - zostanie przygotowany korpus 500 dialogów człowiek-maszyna
  - drugi korpus będzie również anotowany semantycznie po uprzedniej transkrypcji

Zadania dla zespołu języka polskiego

- Zebranie i anotacja korpusu dialogów w dwóch etapach:
  - nagranie 500 dialogów człowiek-człowiek
  - anotacja morfosyntaktyczna i semantyczna dialogów
  - przygotowanie 500 dialogów człowiek-maszyna (najprawdopodobniej Wizard-of-Oz)
  - anotacja morfosyntaktyczna i semantyczna dialogów
- Dziedzina: komunikcja miejska w Warszawie

Warszawska infolinia ZTM

tel. 94-84

Pod ten numer warszawiacy dzwonią średnio 200-300 razy dziennie, aby uzyskać informacje o połączeniach komunikacyjnych, trasach np. autobusów, czy rozkładach jazdy.

Tematyka rozmów:

- Jak dojechać z punktu A do punktu B?
- O której godzinie przyjedzie najbliższy/ostatni/niskopodłogowy autobus?
- Na którym przystanku ktoś powinien wysiąść?
- Gdzie się przesiąść, aby dojechać do A?
- Czy kogoś obejmują zniżki na bilety?
- Dlaczego tramwaj nie przyjechał/spóźnia się?

Zbieranie korpusu dialogów

- Od miesiąca nagrywane są rozmowy z dwóch linii telefonicznych (głos operatora męski i żeński).
- Zebrano już ponad 1500 dialogów. W projekcie zostanie zanotowane 500, pozostały materiał może zostać wykorzystany w innych badaniach.
- Obecnie trwa transkrypcja dialogów.

Kolejne etapy przygotowania korpusu

- Segmentacja tekstu na wypowiedzi
- Transkrypcja dialogów
- Anotacja morfosyntaktyczna
- Anotacja semantyczna:
  - opis dziedziny
  - poziom predykatów
  - koreferencje i anafory
- Akty dialogowe

Dialog – przed ustaleniem zasad transkrypcji

A: dzień dobry przy telefonie . . . , słucham.
B: dzień dobry chciałam się dowiedzieć jak eee mogę się dostać na przystanek euh Legia stadion
A: tak
A: to Legia stadion , czy Szwoleżerów , bo to jest
B: Legia stadion nie <Event desc="b" type="noise" extent="instantaneous" /> albo Szwoleżerów albo <Event desc="*" type="pronounce" extent="next" /> Rozbrat <Event desc="pi" type="pronounce" extent="instantaneous" /> coś tam takie <Comment desc="start of the speaker #1" /> go
B: <Event desc="b" type="noise" extent="begin" /> wszystko <Event desc="b" type="noise" extent="end" /> jedno , bo e <Event desc="b" type="noise" extent="begin" /> chodzi <Event desc="*" type="pronounce" extent="instantaneous" /> mnie <Event desc="b" type="noise" extent="end" /> o to jak gdyby ooo o połączenie zzz

Transkrypcja dialogów 1

Wielkie litery według zwyczajów danego języka, dodatkowo w zapisie akronimów i literowania

Liczby, cyfry zapisane zostaną słowami

Literowanie wyróżnione zostanie wielkimi literami i oznaczone *spelled*: no Tarino comune di Torino [pron=SPELLED-] T O R I N O [-pron=SPELLED]

Akronimy wymawiane jako pełne słowa będą zapisywane wielkimi literami bez odstępów, akronimy literowane – według poprzedniego punktu

Obce słowa dostaną oznaczenie *lang* i etykietę języka: un hôtel à Toulouse avec piscine si possible cet hôtel doit avoir [lang=English-]wellness service[-lang=English], [lang=English-][pron=SPELLED-] C D [-pron=SPELLED] ROM [-lang=English]

Interpunkcja nie powinna być wprowadzana podczas transkrypcji

Transkrypcja dialogów 2

Urwane słowa będą oznaczane przez tyldę (∼): Legia stadion nie albo szwol+[lex=∼] Szwoleżerów albo Rozbrat lub coś tam takiego; a na początku lub na końcu wypowiedzi uzupełniane, gdy to możliwe: [lex=∼-]m[-lex=∼]erci au revoir

Błędna wymowa zostanie poprawiona, ale miejsce będzie zaznaczone: je souhaiterais avoir des [pron=*-]renseignements[-pron=*] sur ma facture

Niezrozumiałe fragmenty oznaczone zostaną dwoma gwiazdkami (**)

Z nakładających się wypowiedzi transkrybowana będzie dominująca

Wahanie, westchnienia itp. dostanie oznaczenie *fil*: [lex=FIL] nie

Zakłócenia otrzymają oznaczenie *noise*: je veux acheter [noise] des actions L'Oréal

Cisza dłuższa niż 1 sekunda będzie oznaczana przez *sil*: rozumiem rozumiem [sil] dobra dobra

Przykład dialogu i transkrypcji

- DIALOG
- TRANSKRYPCJA

Anotacja morfosyntaktyczna 1

Transkrybowany tekst otrzyma dla każdego słowa znaczniki części mowy oraz uzgodnień. Tagset dla każdego języka będzie dostosowany do zaleceń opracowanych w projekcie EAGLES.

Z placu Zamkowego do Wilanowa jedzie autobus sto szesnaście.

```
<words>
<w id="1" word="z" lemma="z" POS="Prep" morph="-" />
<w id="2" word="placu" lemma="plac" POS="Nc"
morph="m3.gen.sg" />
<w id="3" word="Zamkowego" lemma="Zamkowy" POS="ADJp"
morph="masc.gen.sg.pos" >
<w id="6" word="jedzie" lemma="jechać" POS="VV"
morph="3.sg.pres.imperf" />
<w id="7" word="autobus" lemma="autobus" POS="Nc"
morph="m3.nom.sg" />
</words>
```

Anotacja morfosyntaktyczna 2

Anotacja syntaktyczna pogrupuje słowa w podstawowe frazy nominalne i werbalne.

Z placu Zamkowego do Wilanowa jedzie autobus sto szesnaście.

```
<chunks>
<chunk id="1" span="word_1" cat="Prep" />
<chunk id="2" span="word_2..word_3" cat="NP" />
<chunk id="3" span="word_4" cat="Prep" />
<chunk id="4" span="word_5" cat="NP" />
<chunk id="5" span="word_6" cat="VP" />
<chunk id="6" span="word_7..word_ 9" cat="NP" />
</chunks>
```

Anotacja semantyczna

Nazwa projektu zawiera sformułowanie *spoken Language UNderstanding*, a zatem konieczna jest próba opisu znaczenia poszczególnych wypowiedzi.
Anotacja semantyczna podzielona została na następujące trzy poziomy:

- Poziom atrybutów dziedzinowych – obowiązkowy
- Opis struktury predykatów – nieobowiązkowy
- Reprezentacja koreferencji i anafor – nieobowiązkowy

Atrybuty dziedzinowe – anotacja

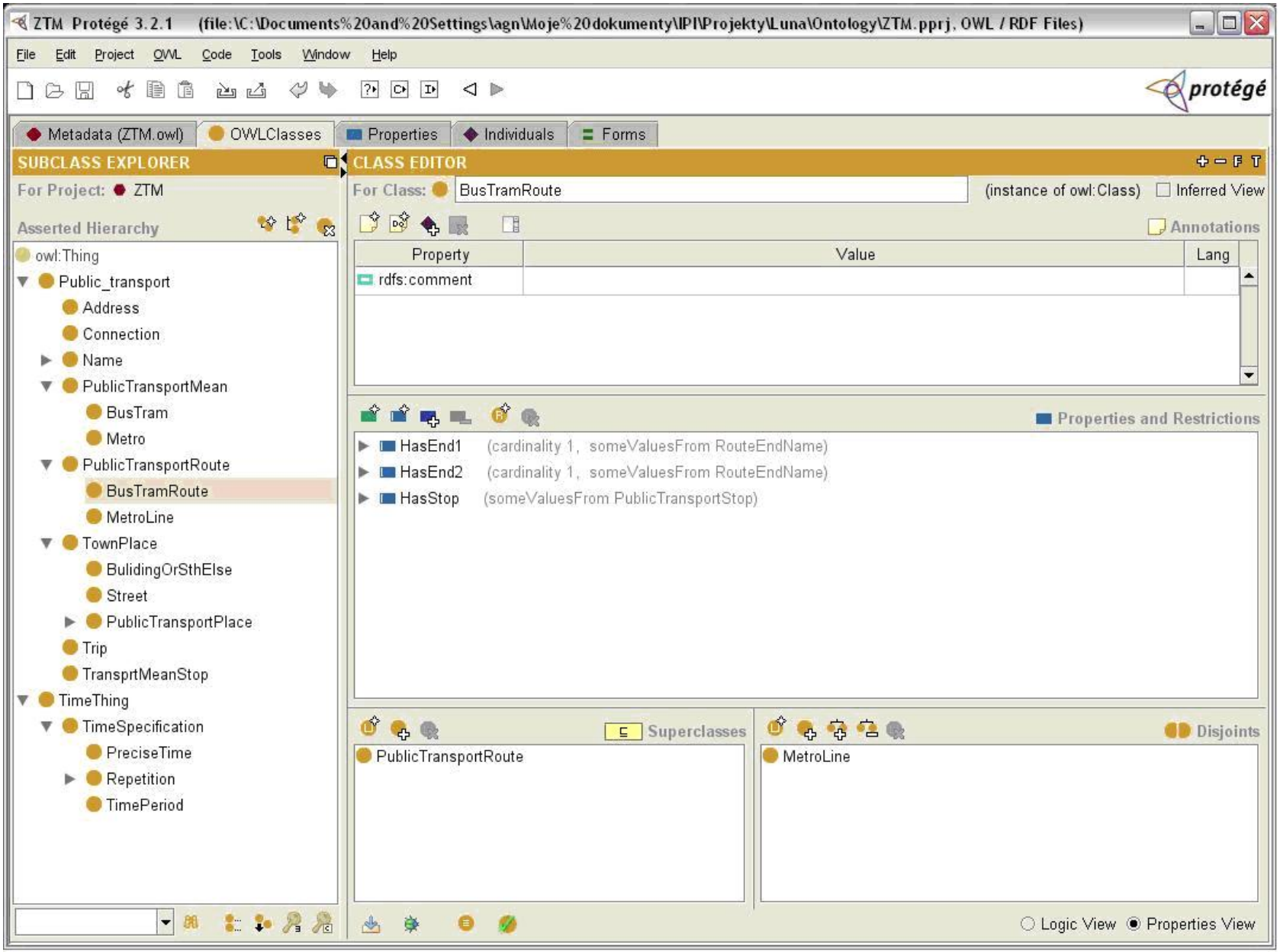
- model dziedziny (standard OWL, edytor Protégé)
- model zawierający hierarchię klas, własności i ograniczenia nałożone na te własności

TownPlace

BuildingOrSthElse

Has BuildingName (some Values from BuildingName)
IsCloseTo (multiple PublicTransportStop)
LocatedAt (some values from StreetName)
HasAddress (cardinality one, some values from Address)

Okno edytora z aktualną wersję ontologii



Atrybuty dziedzinowe – przykład

Z / placu Zamkowego / do / Wilanowa / jedzie / autobus sto szesnaście /

<concept span="word_2..word_3" attribute="streetName" value="plac Zamkowy" />

<concept span="word_5" attribute="buildingOrPlaceName" value="Wilanów" />

<concept span="word_8..word_9" attribute="busLineName" value="sto szesnaście" />

Stuktura predykatów – anotacja

- Słownik czasowników z opisywanej dziedziny
- Koncepcja definicji zgodna z założeniami FRAMENET-u
- Nazwy ról semantycznych przejęte z ontologii dziedziny

Stuktura predykatów – przykład

Frame: DirectTripTaking

Frame-elements: DirectTripStartPoint, DirectTripEndPoint, LineName

Z / placu Zamkowego / do / Wilanowa /
fe_7:DirectTripStartPoint fe_8:DirectTripEndPoint
jedzie / autobus sto szesnaście /
fe_9: target fe_10:LineName

set3={fe_7, fe_8, fe_9, fe_10}

<fe id="fe_7" span="word_2..word_3" frame="DirectTripTaking" frame-element="DirectTripStartPoint" member="set_3" pointer="fe_9" />

<fe id="fe_8" span="word_5" frame="DirectTripTaking" frame-element="DirectTripEndPoint" member="set_3" pointer="fe_9" />

<fe id="fe_10" span ="word_8..word_9" frame="DirectTripTaking" frame-element="LineName" member="set_3" pointer="fe_9" />

Koreferencje – anotacja

- Przyjęty sposób anotacji został zadoptowany z ARRAU (AnaphoRa Resolution And Underspecification), przy czym anotacja ogranicza się do fraz rzeczownikowych odpowiadających obiektom opisanym w modelu dziedziny.
- Główny cel tego poziomu oznaczeń – powiązanie opisów tego samego obiektu w tekście, reprezentacja wiedzy o tym przez etykietę *inf_status*.
- Etykieta *ambiguity* pozwala na określenie niejednoznaczności odwołania (alternatywa obiektów wymienionych na liście *multiple_phrase_antecedent*).
- Etykieta *multiple_phrase_antecedent* pozwala na wprowadzenie listy obiektów, do których odnosi się jedna fraza.
- Etykieta *relation* pozwala na opisanie zależności zachodzącej między anotowanym obiektem, a którymś poprzednio wymienionym, może to być relacja zdefiniowana w modelu dziedziny lub relacja typu *element, podzbiór, nadzbiór*.

Koreferencje – przykłady 1

a [Paris]e1 je vous propose [l'hotel Ibis Montparnasse]e2 et [l'hotel Lafayette]e3

W Paryżu proponuję Hotel Ibis Montparnasse i Hotel Lafayette

<coref id="e_1" span="word_2" inf_status="new" related="no" />

<coref id="e_2" span="word_6..word_8" inf_status="new" related="no" />

<coref id="e_3" span="word_10..word_11" inf_status="new" related="no" />

[ils]e4 ont [une piscine]e5 ?

Czy mają one basen?

<coref id="e_4" span="word_12" inf_status="given" multiple_phrase_antecedent="e_2; e_3" ambiguity="unambiguous" />

<coref id="e_5 span="word_14..word_15" inf_status="new" related="no" />

Koreferencje – przykłady 2

bien, et [cet hotel]e8 accepte [les chiens]e9?

Dobrze, a czy ten hotel akceptuje psy?

<coref id="e_8" span="word_26..word_27" inf_status="given" single_phrase_antecedent="e_3" ambiguity="unambiguous" />

<coref id="e_9" span="word_29..word_30" inf_status="new" related="no" />

[l'hotel Lafayette]e10 n'accepte pas [les animaux]e11

Hotel Lafayette nie akceptuje zwierząt.

<coref id="e_10" span="word_31..word_32" inf_status="given" single_phrase_antecedent="e_3" ambiguity="unambiguous" />

<coref id="e_11" span="word_35..word_36" inf_status="new" related="yes" related_phrase="e_9" relation="superClasseOf" ambiguity="unambiguous" />

Akty dialogowe (Dialogue Acts, DA)

- W projekcie anotacja na tym poziomie nie jest obowiązkowa.
- Oznaczenie aktów potrzebne jest między innymi do tworzenia modelu dialogu lub systemu dialogowego.
- Propozycja wykorzystania schematu DAMSL – Dialogue Act Markup in Several Layers.
- Istnieje bezpłatne narzędzie do tagowania: DAT – Dialogue Annotation Tool (łatwo się zawiesza i wymaga dostosowania).

Akty dialogowe – założenia anotacji

- Wypowiedź (ang. *turn*) może być podzielona na mniejsze fragmenty, którym przypisywane są tagi reprezentujące poszczególne DA. Każdemu fragmentowi można przypisać kilka tagów.
- W projekcie wybrano z DAMSL następujące tagi:
  - Mające wpływ na to co się wydarzy (*Forward looking function*)
    - Statement
    - Action-directive / Open option
    - Committing-speaker-future-action
    - Info-request
  - Stanowiące reakcje na wcześniejsze wypowiedzi (*Backward looking function*)
    - Answer
    - Accept / Reject
    - Signal-understanding / Signal-non-understanding

Akty dialogowe – modyfikacje anotacji

- Czy każdy fragment powinien być otagowany?
- Proponujemy uzupełnienie listy tagów (również z DAMSL):
  - Opening
  - Closing
  - Abandoned (fragmenty wypowiedzi nie na temat)
- Powrót do koncepcji powiązania wypowiedzi stanowiącej reakcję (*Backward looking function*) z fragmentem dialogu zawierającym pytanie lub propozycję.

Przykład

A:	o której jest najbliższy autobus z Dworca Południowego do Powsina ?	(info-request)
B:	o siódmej trzydzieści	(answer)
A:	nie	(reject)
	to może następny?	(info-request)

Aktualny stan prac

- Poziomy anotacji korpusów zostały uzgodnione między partnerami i została przygotowana dokumentacja opisująca (z grubsza) format anotacji.
- Ponad 1500 dialogów z infolinii ZTM jest już nagranych.
- Trwa transkrypcja dialogów, wkrótce rozpocznie się ich anotacja morfosyntaktyczna.
- Jest gotowa wstępna wersja ontologii.
- Pierwsza wersja korpusu dialogów człowiek-człowiek powinna być przygotowana do końca roku.