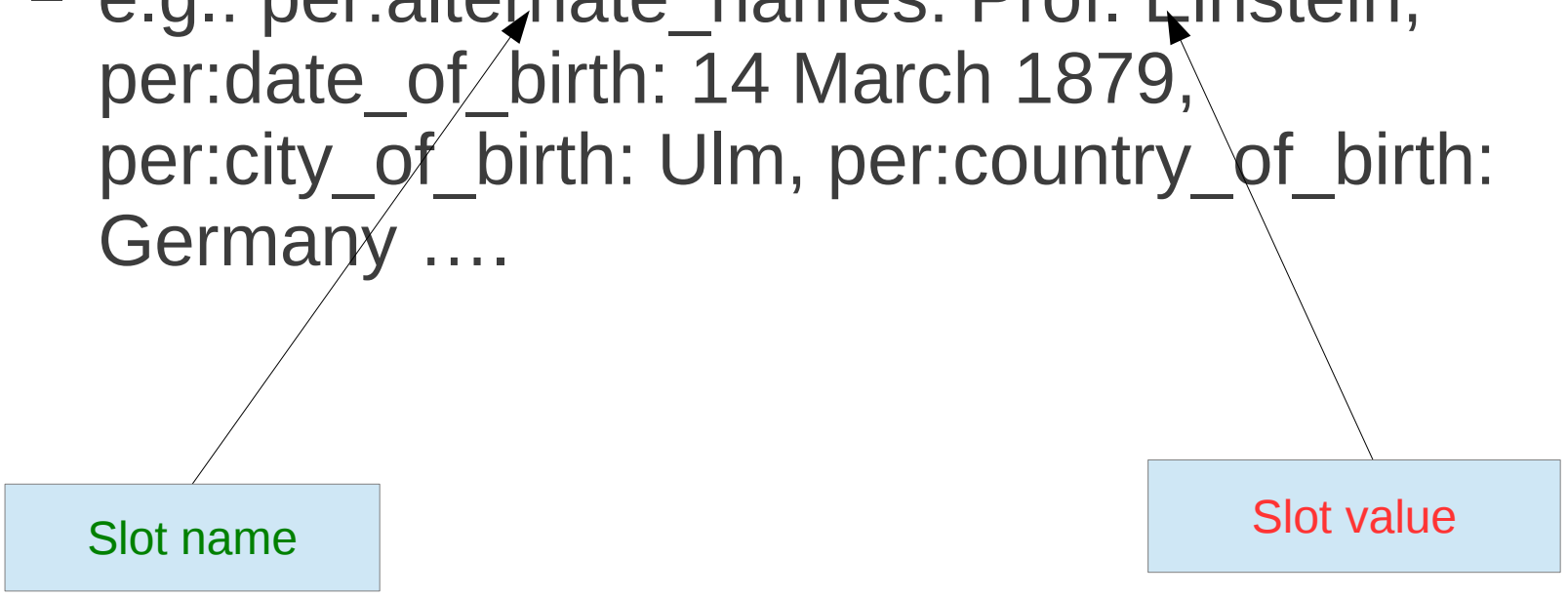
Stanford's Distantly Supervised Slot Filling System

- Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, Christopher D. Manning. *Stanford's Distantly-Supervised Slot-Filling System.* **Proceedings of the TAC-KBP 2011 Workshop, 2011.**
- Mihai Surdeanu, David McClosky, Julie Tibshirani, John Bauer, Angel X. Chang, Valentin I. Spitkovsky, Christopher D. Manning. *A Simple Distant Supervision Approach for the TAC-KBP Slot Filling Task.* **Proceedings of the TAC-KBP 2010 Workshop, 2010**
- `http://nlp.stanford.edu/software/mimlre.shtml`

KBP task definition

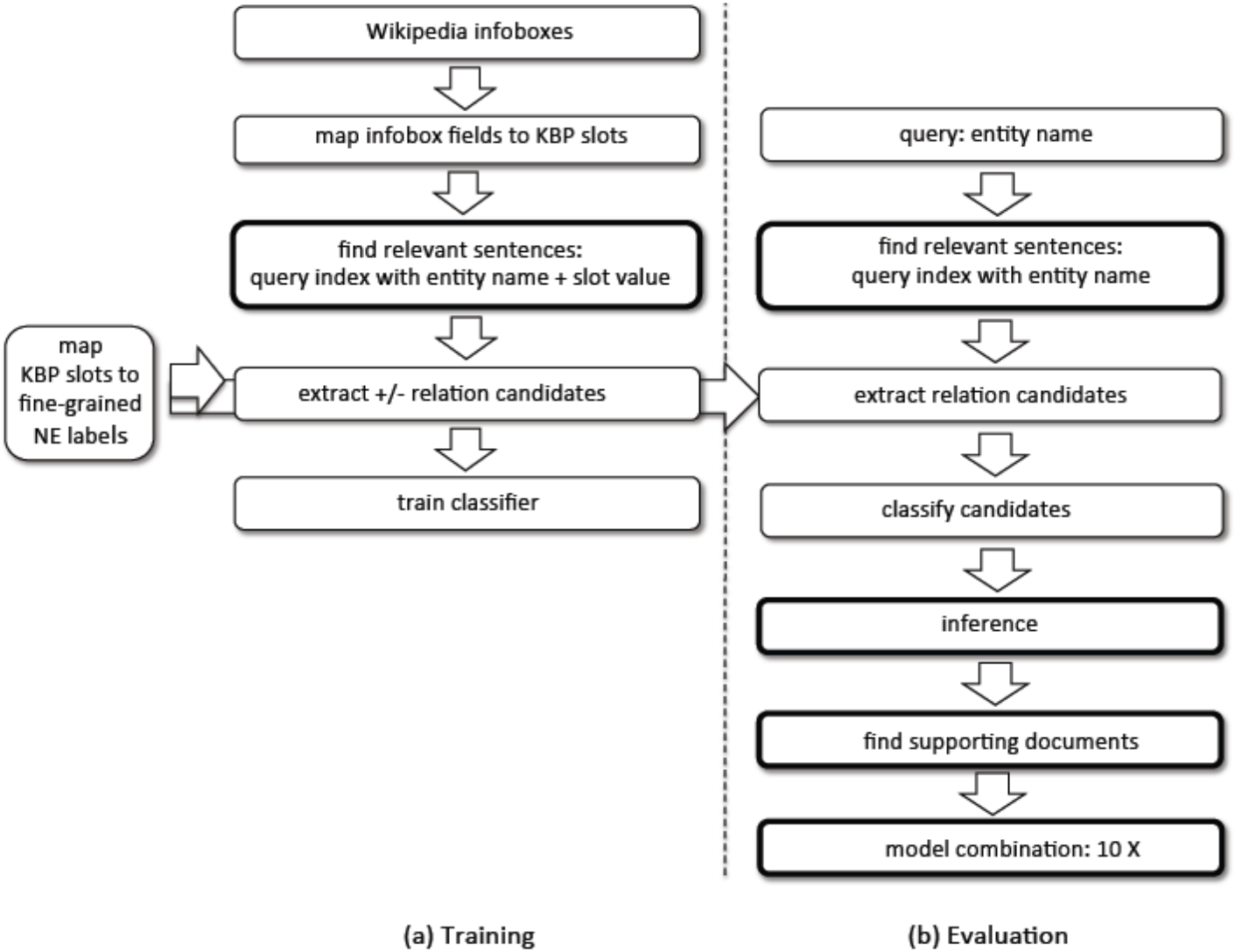
- Input : name of entity, type(PER/ORG)
  - e.g.: Albert Einstein, PER
- Output : Attributes of this slot
  - e.g.: per:alternate_names: Prof. Einstein, per:date_of_birth: 14 March 1879, per:city_of_birth: Ulm, per:country_of_birth: Germany ….



Wikipedia info-boxes

- <entity wiki_title="Mack_McLarty" type="PER" id="E0004232" name="**Mack McLarty**">
- <facts class="Infobox US Cabinet official">
- <fact name="name">Thomas F. "Mack" McLarty III</fact>
- <fact name="order">17 th</fact>
- <fact name="title"><link>White House Chief of Staff</link></fact>
- <fact name="term_start"><link>1993</link></fact>
- <fact name="term_end">January, <link>1994</link></fact>
- <fact name="predecessor"><link entity_id="E0214786">James Baker</link></fact>
- <fact name="successor"><link entity_id="E0076158">Leon Panetta</link></fact>
- <fact name="president"><link entity_id="E0100324">Bill Clinton</link></fact>
- <fact name="birth_date">June 14, 1946 (1946-06-14) (age 62)</fact>
- <fact name="birth_place"><link entity_id="E0600134">Hope</link>, <link entity_id="E0240539">Arkansas</link></fact>
- <fact name="party"><link>Democratic</link></fact>
- </facts>
- <wiki_text><![CDATA[Mack McLarty
-
- Thomas F. "Mack" McLarty III (born June 14, 1946) is a prominent Arkansas
- business and political leader and former White House Chief of Staff for US
- President Bill Clinton. He is the President of McLarty Associates (his
- Washington-based consulting company - see www.maglobal.com), as well as Chief
- Executive Officer of the McLarty Companies.
- ……..
- ……..
-]]></wiki_text>
- </entity>
-
-

System Architecture



KB file

Features

<entity wiki_title="Asha_Puthli" type="UKN" id="E0007544" name="Asha Puthli">
<facts class="Infobox musical artist">
<fact name="Name">Asha Puthli</fact>
<fact name="Background">solo_singer</fact>
<fact name="Origin"><link>Bombay</link>, <link entity_id="E0031352">India</link></fact>
<fact name="Genre"><link>Jazz</link>,
<link>soul</link>, <link entity_id="E0783192">funk</link>,
<link entity_id="E0668346">pop</link>,
<link entity_id="E0390417">ambient music</link>,
<link entity_id="E0539140">electronica</link>, <link>Indian music</link></fact>
<fact name="Years_active">1970—present</fact>
<fact name="Label">CBS / Sony, Polygram, TK Records, Autobahn Records, Top of the World</fact>

- Relation Features used with examples:
  - arg_type : arg1type=ENT:PERSON_and_arg2type=COUNTRY
  - arg_order :
  - full_tree_path : NNP_<-_NP_<-_S_<-_VP_<-_NP_->_NP_->_NNP
  - surface_distance_binary : surface_distance_6
  - surface_distance_bins : surface_distance_bin_lt10
  - adjacent_words : leftarg0-diva rightarg0-, leftarg1-featuring rightarg1--
  - entities_between_args : entity_between_args:_MODIFIER
  - entity_counts_binary : entity_counts_COUNTRY:_1.0, entity_counts_ENT:PERSON:_1.0, entity_counts_MODIFIER:_1.0, entity_counts_NUMBER:_1.0
  - entity_counts_bins : entity_counts_bin1 entity_counts_bin1 entity_counts_bin1 entity_counts_bin1
  - span_words_unigrams : span_word:jazz span_word:diva
  - dependency_path_lowlevel : _dep->_partmod->_
  - dependency_path_words : word_in_dependency_path:bear

Datum file

Implementation details

- Sentences are stored in an inverted index which are queried using "*entity_name* + *slot value*"
- The code to create index file is part of the package
  - Input: corpus of documents
  - Output: Lucene inverted index of the sentences
- The index file is read along with the KB file (slot information) to create datum files
  - Pre-processed training data files used as input to the classifiers