Aleksandar Savkov
EuroMatrixPlus Project
Linguistic Modelling Laboratory
Bulgarian Academy of Sciences

Integration of Bulgarian Web Services

D-SPIN Workshop Freudenstadt

16. November 2010

- Providing access to Language Resources
  - Concordance over text and annotated corpora (morphological, syntactic, etc)
- Providing access to Language Technology
  - Tokenizer & Sentence Boundary Detector
  - POS Tagger
  - Lemmatizer
  - Parser

Bulgarian Web Services

- Providing access to Language Resources
  - Concordance over text and annotated corpora (morphological, syntactic, etc)
- Providing access to Language Technology
  - Tokenizer & Sentence Boundary Detector
  - POS Tagger
  - Lemmatizer
  - Parser

Bulgarian Web Services

Corpus query tool

- Provides access to the BulTreeBank: Bulgarian Reference Corpus
  - allows the user to search the Bulgarian Reference Corpus using regular expressions
  - provides the context of the queried expression in alphabetical order
  - and term frequencies in the cases where regular expressions were used
  - currently it is used for concordance related searches

- Provides access to the BulTreeBank: Bulgarian Reference Corpus
  - allows the user to search the Bulgarian Reference Corpus using regular expressions
  - provides the context of the queried expression in alphabetical order
  - and term frequencies in the cases where regular expressions were used
  - currently it is used for concordance related searches

Corpus query tool

Corpus Web Service

- Current state:
  - Lucine indexing and document search
  - CLaRK regular expressions implementation
  - a web GUI is available at http://www.webclark.org
- Future development:
  - additional context ordering based on metrics of context similarity
  - TCF integrated web service
    - + beneficial for automated processing of concordance results
    - + needs new TCF tags to be implemented

- Current state:
  - Lucine indexing and document search
  - CLaRK regular expressions implementation
  - a web GUI is available at http://www.webclark.org

- Future development:
  - additional context ordering based on metrics of context similarity
  - TCF integrated web service
    - beneficial for automated processing of concordance results
    - needs new TCF tags to be implemented

Corpus Web Service

- Tokenizer & Sentence Boundary Detector
- POS Tagger
- Lemmatizer
- Parser

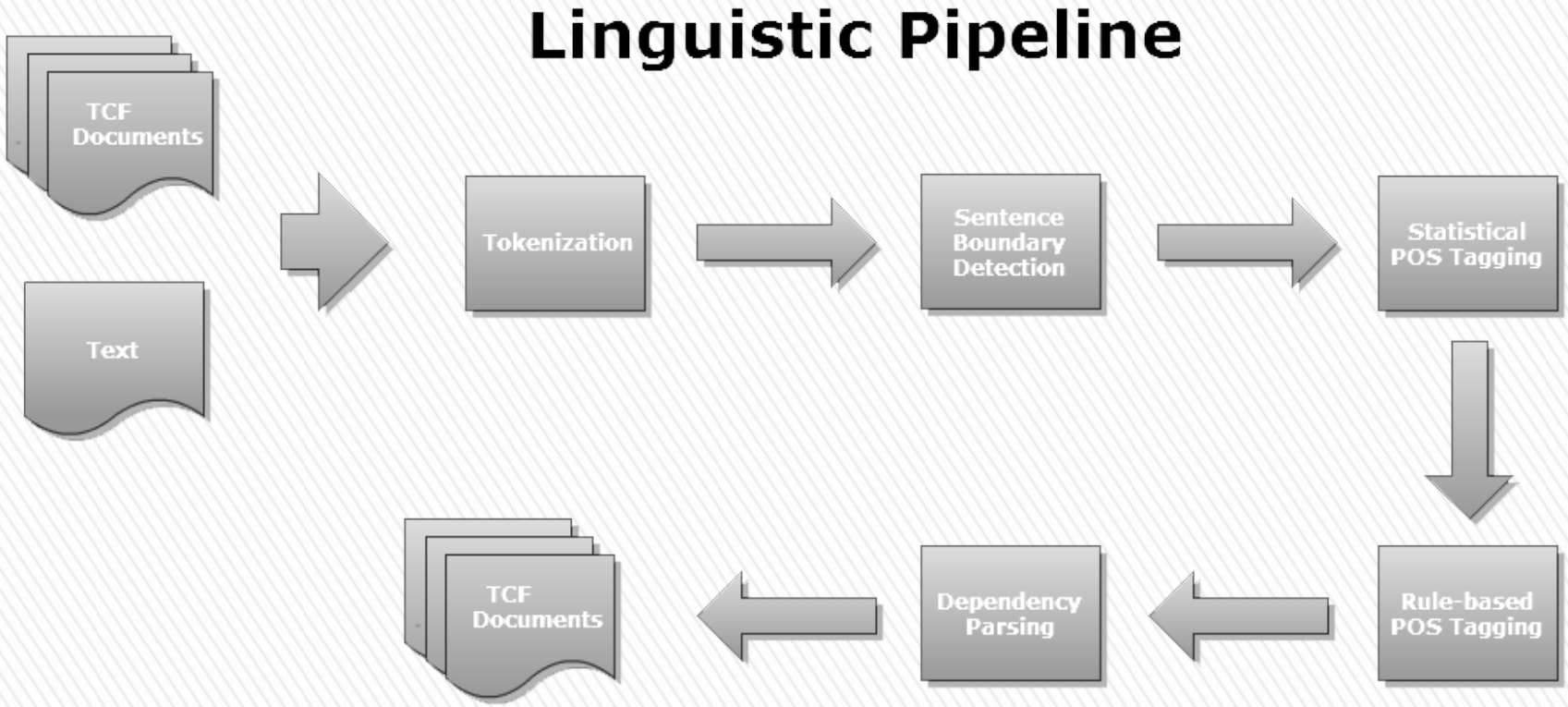
- Most of them will be used in the development of the EuroMatrixPlus Project pipeline
  - the aim of the project is to create an English-Bulgarian parallel treebank needed for statistical machine translation

Language Technology Web Services

- Tokenizer & Sentence Boundary Detector
- POS Tagger
- Lemmatizer
- Parser

- Most of them will be used in the development of the EuroMatrixPlus Project pipeline
  - one of the aims of the project is to create an English-Bulgarian parallel treebank needed for statistical machine translation

Language Technology Web Services



Linguistic pipeline

Tokenizer and Sentence Boundary Detector

- Naïve implementations:
  - regular expression based tokenizer
  - simple punctuation-based sentence detector aided by acronyms, abbreviations, and names dictionaries.
- Finite-state grammars implementations:
  - will be implemented in CLaRK
  - using the same linguistic resources acronyms, abbreviations, and names dictionaries
- Web services:
  - currently: naïve back-end with a web GUI
  - working on: TCF integration

Tokenizer and Sentence Boundary Detector

- Naïve implementations:
  - regular expression based tokenizer
  - simple punctuation-based sentence detector aided by acronyms, abbreviations, and names dictionaries.
- Finite-state grammars implementations:
  - will be implemented in CLaRK
  - using the same linguistic resources acronyms, abbreviations, and names dictionaries
- Web services:
  - currently: naïve back-end with a web GUI
  - working on: TCF integration

- Naïve implementations:
  - regular expression based tokenizer
  - simple punctuation-based sentence detector aided by acronyms, abbreviations, and names dictionaries.

- Finite-state grammars implementations:
  - will be implemented in CLaRK
  - using the same linguistic resources acronyms, abbreviations, and names dictionaries

- Web services:
  - currently: naïve back-end with a web GUI
  - working on: TCF integration

Tokenizer and Sentence Boundary Detector

Statistical POS Tagging

- SVMTool
  - Support Vector Machine (SVM) based tagger developed by Jesús Giménez and Lluís Márquez
  - Based on Vapnik's SVM implementaition *SVM Light* by Thorsten Joachims
  - Trained on a tagged corpus part of the BulTreeBank corpus (of around 296K tokens) using additional dictionary (1 million word forms)
- Results:
  - 93.5470% accuracy tested on a 12K homogenous corpus
  - 90.4710% accuracy tested on a 57K fiction corpus
- Working on:
  - rules to fix recurring tagger errors

Statistical POS Tagging

- SVMTool
  - Support Vector Machine (SVM) based tagger developed by Jesús Giménez and Lluís Márquez
  - Based on Vapnik's SVM implementaition *SVM Light* by Thorsten Joachims
  - Trained on a tagged corpus part of the BulTreeBank corpus (of around 296K tokens) using additional dictionary (1 million word forms)
- Results:
  - 93.5470% accuracy tested on a 12K homogenous corpus
  - 90.4710% accuracy tested on a 57K fiction corpus
- Working on:
  - rules to fix recurring tagger errors

Statistical POS Tagging

- SVMTool
  - Support Vector Machine (SVM) based tagger developed by Jesús Giménez and Lluís Márquez
  - Based on Vapnik's SVM implementaition *SVM Light* by Thorsten Joachims
  - Trained on a tagged corpus part of the BulTreeBank corpus (of around 296K tokens) using additional dictionary (1 million word forms)
- Results:
  - 93.5470% accuracy tested on a 12K homogenous corpus
  - 90.4710% accuracy tested on a 57K fiction corpus
- Working on:
  - rules to fix recurring tagger errors

POS Tagging Web Service

- Current state:
  - testing a web GUI service that provides Tokenization, Sentence Boundary Detection and statistical POS Tagging
  - the back-end is a Java implementation wrapping the (currently naïve) Tokenization and SBD, and the Perl implementation of SVMTool
- Working on:
  - TCF input and output capabilities for each state of the current services' pipeline

POS Tagging Web Service

- Current state:
  - testing a web GUI service that provides Tokenization, Sentence Boundary Detection and statistical POS Tagging
  - the back-end is a Java implementation wrapping the (currently naïve) Tokenization and SBD, and the Perl implementation of SVMTool

- Working on:
  - TCF input and output capabilities for each state of the current services' pipeline

- Available tools:
  - Rule-based POS Tagging Module implemented through constraints within CLaRK
  - Lemmatizer for Bulgarian with results going over 98% when applied on POS tagged input
  - MaltParser (dependancy parser) developed by Svetoslav Marinov and trained on the BTB corpus

Future Development

Integration of Web Services

- Integration with WebLicht
  - all services will be equipped with TCF I/O capabilities
  - and contributed to the WebLicht project
- BTB Group internal integration
  - provide access to whole pipelines
  - develop an internal format to enhance the speed of data transfer between services
  - accept plain text and TCF input
  - output TCF

- Integration with WebLicht
  - all services will be equipped with TCF I/O capabilities
  - and contributed to the WebLicht project

- BTB Group internal integration
  - provide access to whole pipelines
  - develop an internal format to enhance the speed of data transfer between services
  - accept plain text and TCF input
  - output TCF

Integration of Web Services

» Questions?

Thank you!