WP6 – Lexical Acquisition

UCAM, UPF, CNR-ILC, ILSP

Objectives

- Development of techniques for automatic acquisition of
  - subcategorization frames
  - selectional preferences
  - multiword expressions
  - lexical-semantic classes
- Starting point: a comprehensive analysis of existing techniques for different languages

Objective

- Building on the best existing techniques, improve their
  - accuracy
  - scalability
  - portability between domains
- Use them to extract monolingual and domain-specific lexica from suitably annotated corpora
- Build a component which merges automatically acquired lexicons with existing dictionaries. The resulting component will be included in the platform.

Description

WP6.1	Methods for subcategorization, selectional preference and multiword acquisition
WP6.2	Lexical-semantic classification methods
WP6.3	Merging of dictionaries

Deliverables

D6.1 (t6) Report on technologies / tools to be developed & integrated, evaluation criteria, resource specification

D6.2 (t28-30) Integrated final lexical acquisition components, technical description (scientific paper)

D6.3 (t28-30) Monolingual lexicons, tuned to a chosen domain using acquisition techniques

D6.4 (t28-30) Lexical merger

D6.5 (t28-30) Merged dictionary

Internal deliverables (t13, t21, t29)

Components will be integrated in the platform (WP3)

Work plan for the first 6 months

Integrated report (covering the different languages) which includes

- survey of the state of the art
- survey of existing tools (or relevant resources)
- work plan: tools to be developed and integrated
- evaluation
- resources to be produced
- detailed work plan for the rest of the project

Tools available to partners

	English	Spanish	Italian	Greek	French
SCF dict.	**yes**	**yes**	**yes**	**yes**	**yes**
Wordnet	**yes**	**yes**	**yes**	**yes**	**yes**
VerbNet	**yes**	**yes**			
Chunker	n/a	n/a	n/a	n/a	n/a
Treebank	n/a	n/a	n/a	n/a	n/a
Shallow parser	**yes**	**yes**	**yes**	**yes**	
SCF	**yes**				**yes**
SP	**yes**				**yes**
Lex. classes	**yes (verbs)**	**yes (nouns)**			**yes (verbs)**
MWE					

Focus

Subcategorization	John broke the window **NP**
Selectional preferences	**HUMAN** break **OBJECT**
Lexical-semantic classes	John **BREAK** window
Multiword expressions	John and Mary **broke up**

- Key questions:
  - which type(s) of information will benefit MT the most?
  - what is realistic for each language? (given existing resources and tools, licensing issues, the time available...)
  - what kind of resources will we acquire? (which lex info, general / domain, which languages)
  - lexical merger (ditto...)

Subcategorization

Goal: a system which learns subcategorization frame types and frequencies for verbs from corpora (plus gathers information about verb tense, voice, subjects, objects, etc.)

break

- John broke the window **NP** 0.35
- John broke the window with a hammer **NP + PP** 0.20
- The window broke **INTRANS** 0.15
- etc.

SCF acquisition

- **Required resources:**
  - corpora (min. 100 occurrences per verb)
  - tagger, tokeniser, lemmatizer, shallow parser (pref. dependency) which does not use SCFs during parsing (or at least a chunker, or the last resort: a treebank)
  - SCF dictionaries for development & evaluation

- **Effort:**
  - A classifier which matches parses / grammatical relations with SCF types
  - A lexical builder which constructs SCF entries from classified data
  - A filter which removes noisy SCFs
  - Evaluation of SCFs against dictionaries & manual analysis

Previous research

- **Spanish:** Pazos et. al. 2009. Semi-automatic Generation of Subcategorization Frames for Spanish Verbs Using Ontologies and Verbs Functional Class. Jnl of Computers. 4(8). 721-727.

- **Italian:** D. Ienco, S. Villata and C. Bosco. 2008. Automatic extraction of subcategorization frames for Italian. In Proc. of LREC. *Marrakech, Morocco.*

- **Greek:** Kermanidis et al., 2004. Automatic acquisition of verb subcategorization information by exploiting mininal linguistic resources. Int. Jnl. of Corpus Linguistics .

Selectional preferences

Goal: a system which identifies semantic classes of nous appearing as arguments of verbs

break

- **John** broke **the window** with **a hammer**

- **AGENT / HUMAN** (John, he, Mary, ...)
- **BREAKABLE OBJECT** (window, glass, vase...)
- **INSTRUMENT** (hammer, stick, hand...)

Selectional preferences

- **Required resources:**
  - corpora
  - tagger, tokeniser, lemmatizer, parser (optionally: a SCF system)
  - (optionally: WordNet –style resources)

- **Effort**
  - One easy method: cluster nouns appearing in argument slots of verbs in parsed data
  - Extract features from dependency relations, cluster them using a suitable method
  - Evaluate manually / against lexical resources / in the context of a task (e.g. pseudo disambiguation)

Lexical-semantic classes

Goal: a system which identifies lexical (syntactic-semantic / semantic) classes of verbs in corpora

Verbs

- **BREAK:** break, fracture, rip, smash...
- **PUT:** put, place, position, fill....

Nouns

- **COUNT** vs. **MASS** nouns, etc.

Lexical-sem. verb classes

- **Required resources:**
  - Corpora
  - Tagger, Tokeniser, Lemmatizer, (Parser, a SCF system)
  - VerbNet / WordNet –style resources for evaluation

- **Effort:**
  - Extract a range of lexical, (syntactic, semantic features) (shallow or deep) from corpora
  - Cluster / classify words using machine learning
  - Evaluate against existing classes and/or manually,

Multiword Expressions

- **Required resources:**
  - Corpora
  - Tagger, Tokeniser, Lemmatizer, (Parser, a SCF system)
  - Dictionaries for development and evaluation

- **Effort depends entirely on MWEs one wants to focus on, e.g.**
  - Collocations (salt - pepper)
  - Compound nouns (a google map)
  - Verb particle constructions (break up)
  - Light verbs (take a walk)
  - Idioms (spill the beans)
  - etc. etc. etc.

Plan

- Languages
  - English, Spanish, Italian, Greek, (possibly French)
- The choice of acquisition techniques:
  - Priority area for MT? (SCF acquisition?)
  - What is realistic for the language in question (given existing resources, tools, licenses, and the time available)?
  - What do we want to do?
- Lexicons
  - What kind of resources are we going to build and for which languages (licensing issues)
- Lexical merger
  - (as above...)

Plan

		English	Spanish	Italian	Greek	French
Types of lexical info	**SCF**	**yes**	**yes**	**yes**	**yes**	
	SP	**yes**	**yes**	**yes**		
	CLASSES verbs nouns	**yes**	**yes**	**?** **?**		
	MWE		**yes**	**yes**		
Lexicons	SCF SP Lex class	**yes** **yes**				
Lexicon types	general domains	**yes** **yes**				
Merger		**?**		**?**		

The first deliverable (month 6)

Integrated report (covering the different languages) which includes

- survey of the state of the art
- survey of existing tools (or relevant resources)
- work plan: which tools will we develop and how
- evaluation
- resources to be produced
- detailed work plan for the rest of the project