Bootstrapping Language-neutral Term Extraction

Wauter Bosma <w.bosma@let.vu.nl>
Piek Vossen <p.vossen@let.vu.nl>

The KYOTO Project

- Partners across Europe & Asia:
  - Technical: EHU, CNR, NICT, VUA, AS, BBAW, MUNI, Synthema, Irion;
  - Users: WWF, ECNC;
- 7 languages (Basque, Chinese, Dutch, English, Italian, Japanese, Spanish);
- Website: www.kyoto-project.eu

The KYOTO Knowledge Cycle

Semantics in Text

- Goal: domain modelling (*facts* & *concepts*)
- Example: *terrestrial species declined by 55%*
- Terms are components of facts:
  - Decline
  - 55%
  - Terrestrial species

Term Extraction

- Identify domain terms (ranked list);
- **Identify term relations**;
- Example:
  - *Terrestrial species* $\subset$ *species*
  - *Terrestrial species* $\cap$ *marine species* = $\varnothing$
  - *Frog* $\in$ *amphibious species*

Strategies of Automatic Term & Relation Extraction

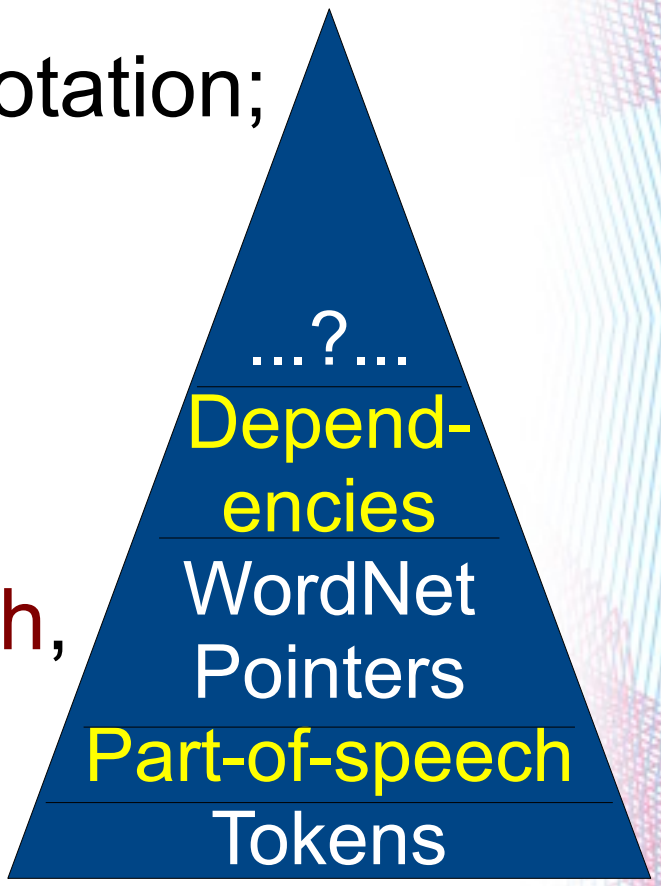
- Morpho-syntactic analysis (e.g., *terrestrial species* ⊂ *species*);
- Pattern-based analysis (e.g., *amphibious species such as frogs*);
- Distributional statistics (terms used similarly are similar);
- Language alignment by means of wordnet mappings;
- Our strategy: use a combination of the above for extracting relations and ranking terms.

Term & Relation Extraction in KYOTO

- Pre-processing: part-of-speech, dependencies, word sense disambiguation;
- Extract (plenty of) candidate terms;
- Extract relations using a combination of methods (morpho-syntactic, pattern-based, distributional, language alignment);
- Use relations and document frequencies to rank terms for domain-relevance.

Step 0: Pre-processing

- KAF – KYOTO Annotation Format;
- Supports arbitrary layers of annotation;
- Extendible;
- Language-neutral;
- Used with KYOTO languages:
  Basque, Chinese, Dutch, English, Italian, Japanese, Spanish;
- KAF is our starting point for term extraction.



KAF

Term Database

- Terms (including features such as domain-relevance, part-of-speech, etc.);
- Relation types (including features such as transitivity, commutativity, etc.);
- Internal relations (between terms);
- External relations (between a term and a resource such as WordNet);
- Term instances (with pointer to source).

Step 1: Candidate Terms

- Nouns (or other POS) are candidate terms (e.g., *species*);
- The head of compound nouns are candidate terms (e.g. *landbouwbeleid*, *beleid*);
- Noun phrases are candidate terms (e.g., *vertebrate terrestrial species*);
- Reduced noun phrases are candidate terms. Modifiers are stripped one by one, towards the head:
  - vertebrate terrestrial species ⇢ terrestrial species ⇢ species
  - migration of species ⇢ migration of ⇢ migration

Step 2: Morpho-syntactic Analysis

- A noun phrase is a hyponym of derived reduced noun phrases (e.g., *terrestrial species* ⊂ *species*);
- A compound is a hyponym of its head (e.g., *landbouwbeleid* ⊂ *beleid* – *agricultural policy* ⊂ *policy*).

Step 3: Pattern-based Analysis

- Learning patterns from existing resources, eg. wordnets, species2000.
- Wordnet: hyponym(frog,amphibian)
- Corpus: *… amphibians such as frogs ...*
- Pattern: X such as Y
- Corpus: *… habitat for wading birds such as golden plover, lapwing and redshank*;
- Corpus: *Notable trends include the recent recovery of the pinkfooted goose, avocet and ...*

Enumerations

- … *golden plover*, *lapwing* and *redshank*.
- *… limiting the use of fertilisers, manures and pesticides;*
- Share a syntactic function;
- Share a common hypernym or attribute;
- Usually disjoint (*LREC attracted over 1000 researchers and people*);

Step 4: Distributional Statistics

- "Terms used in a similar way are similar";
- Measure the amount of shared context;
- Context can be anything, e.g.: linear context, dependency relations, etc.
- High similarity statistic is evidence of a shared hypernym or attribute.

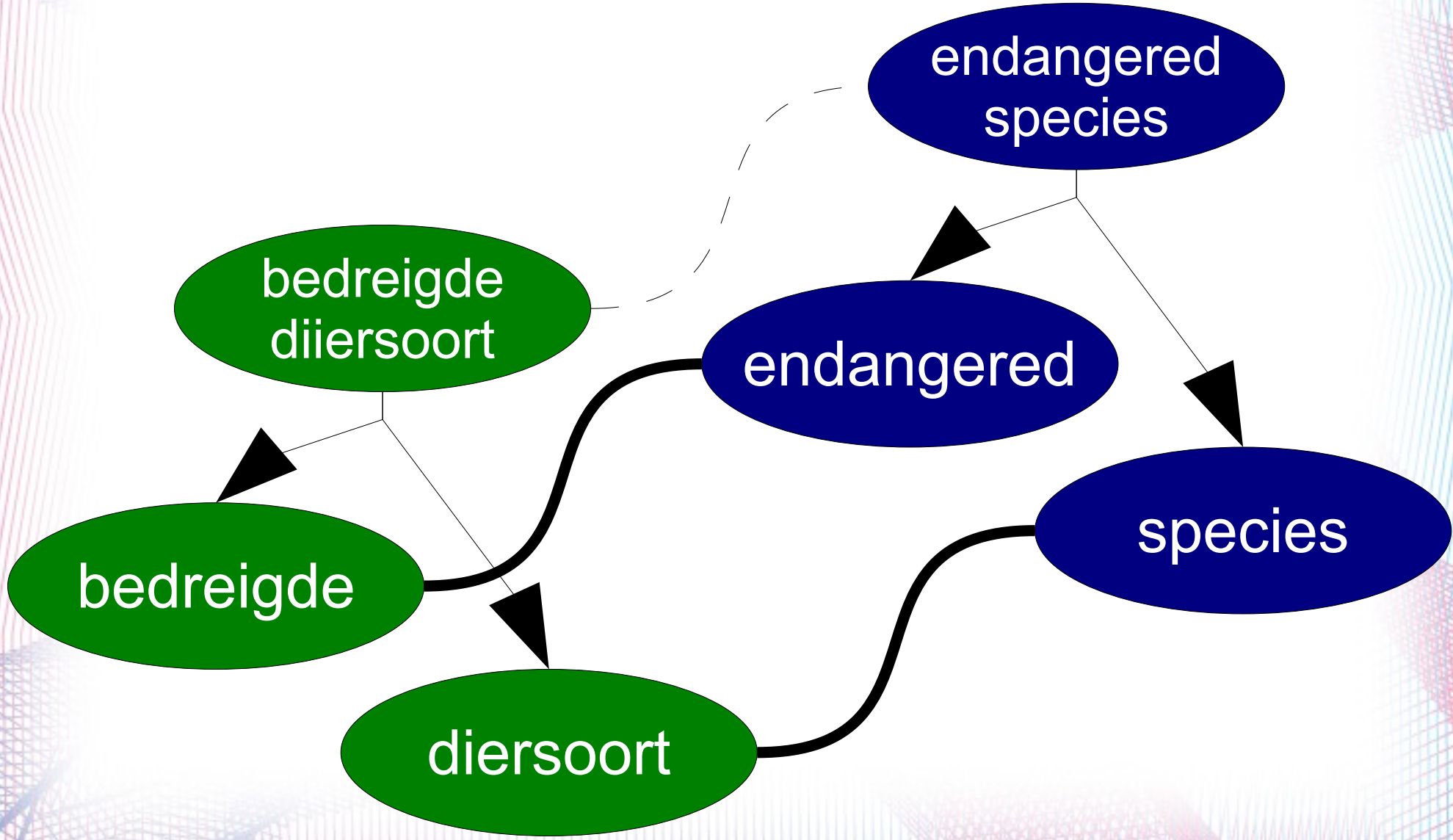
Step 5: Ranking Terms

- Distinguish domain-relevant terms from non-terms;
- (As opposed to distinguishing domain terms from generic terms;)
- No clear boundary;
- A confidence value is assigned to each candidate term, representing its 'termness';
- The confidence value is calculated from the term relation graph and occurrence frequency;
- Candidate terms above a certain confidence threshold may be regarded terms.

Step 6: Language Alignment

- Wordnet mappings provide relations between languages;
- Wordnets, term database and other resources provide relations within a language;
- Infer new relations between languages;

Language Alignment: Example



Evaluation

- Gold standard for evaluation must be
  - corpus-based;
  - exhaustive.
- No such resource exists;
- We need to create one.

Conclusion

- Based on language-neutral KAF;
- Term relations to leverage term ranking;
- Domain terms may improve parsing;
- Works with 7 KYOTO languages;