Resolving Orphaned Parts in Taxonomic Descriptions with Machine Learning and Natural Language Processing Methods

Steven S. Chong[1,2], Dongfang Xu[1], Thomas Rodenhausen[1], Hong Cui[1*]

1. University of Arizona, School of Information, Tucson, AZ, USA 2. National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, CA, USA

*hongcui@email.arizona.edu

Introduction

Biodiversity literature contains vast amounts of information in human-readable formats. Morphological descriptions can be parsed to extract data for biological research.

Problem: Descriptions often contain non-specific structural parts (e.g. surface, apex, tip) not explicitly linked to their respective anchor organs. Bridging non-specific structures with anchors is necessary for machines to extract character information.

We compared different methods for resolving meronym (part-of) relations between non-specific parts and anchor organs.

Goal: Associate non-specific structure terms with their anchors because resolving part-of relationships is needed to correctly extract phenotypic characters

Task Example

Example description:

> **Leaflets** *articulated, inserted near the edges of the* **rhachis** towards the *adaxial side, lacking a differently coloured basal gland; stomata on lower surface only or on both surfaces; epidermal cells elongated parallel to long axes of* **leaflets.**

Non-specific structure terms	Anchor (parent) terms
1. edges	1. rhachis
2. adaxial side	2. leaflets
3. lower surface	3. leaflets
4. surfaces	4. leaflets
5. axes	5. leaflets

Data

- Corpus: 3876 descriptions (7562 sentences) covering 11 taxon groups
- Example data sources: Plazi.org, Flora of North America
- Domain experts identified 39 non-specific structures
- Development dataset to develop the two relation identification methods (169 sentences, random sample)
- Test dataset to expand taxon and non-specific structures coverages (167 sentences, stratified-random sample)

Methods

Preprocessing

- Explorer of Taxon Concepts (ETC; Cui et al., 2016) Toolkit used to annotate structures, characters, and relationships in both development and test data as input for algorithms
- Created ontologies to indicate part-of relationships between structure terms in development and test data

Relation Identification Methods

1) **Syntactic rules**:
   - Candidate anchor organs located within three-sentence boundary of non-specific structure terms
   - Part-of relationships from ETC Toolkit involving “of-phrases” (e.g. blades of the leaves)
   - Possession words around a non-specific structure term
   - The non-specific structure ontology

2) **Support vector machine (SVM)**:
   Pairwise Classification
   - For each anchor term, classify binary relations for all candidate non-specific structure terms and select those with highest probabilities

   Feature Groups
   1. Distance and position features
   2. Bag-of-word features (e.g. “in”, “on”, “contains” before/after structure terms)
   3. Semantic features from the ontology

Results

Two baseline algorithms were implemented for comparison purposes:

- Baseline 1 chose subject entity in a sentence as its anchor term
- Baseline 2 selected nearest entity term to non-specific structure as its anchor term

Precision (P), recall (R), and F1 scores were calculated for the test and development datasets.

Table 1: Performances of the Two Methods and Baseline Algorithms

Methods	F1 (Development)	P (Test)	R (Test)	F1 (Test)
Baseline 1 (subject entity)	63.9%	42.3%	42.3%	42.3%
Baseline 2 (closest entity)	30.3%	33.2%	33.2%	33.2%
Syntactic (ontology only)	91.1%	92.2%	90.5%	91.4%
Syntactic (all rules)	93.7%	93.0%	91.3%	92.1%
SVM (feature groups 1 and 2)	76.1%	60.9%	60.9%	60.9%
SVM (all features)	89.6%	80.7%	80.7%	80.7%

Of the 366 non-specific structure term occurrences in the test dataset:

- SVM incorrect in 58 cases
- Syntactic method incorrect in 25 cases
- Both SVM and syntactic methods incorrect in 7 cases

Web Resources

Explorer of Taxon Concepts (ETC) Toolkit:
http://etc.cs.umb.edu/etcsite/
Syntactic method source code:
https://github.com/biosemantics/charaparser/tree/master/enhance
SVM source code:
https://github.com/biosemantics/SVM-for-Nonspecific-Structure

Conclusions

- Ontologies were reliable knowledge sources for resolving orphaned parts in morphological descriptions.
- The results of the syntactic and SVM methods were complementary and mistakes rarely overlapped.
- The syntactic method performed better than the SVM method and will be implemented in the ETC Toolkit… but future research will examine the complementary nature of both methods.

Acknowledgements

This material is based upon work supported by the U.S. National Science Foundation under Grant No. DBI-1147266.

Selected Reference

Cui H, Xu D, Chong SS, Ramirez M, Rodenhausen T, Macklin JA, Ludäscher B, Morris R, Soto E, Koch NM (2016) Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. BMC Bioinformatics 17 (1): 471. https://doi.org/10.1186/s12859-016-1352-7