Automated Learning Terminological Ontologies
Problem
- Concept hierarchies are widely (e.g., in digital libraries) for purposes of navigation, browsing, query suggestion and document retrieval.
- They are traditionally designed and maintained manually.
- Manual approach is time-consuming and prone to obsolescence.

Objective
To learn concept hierarchies and terminological ontology from unstructured text.
Plausibility of learning is based on the assumption “given sufficient large amount of text in a domain, coverage of knowledge in that domain can be ensured”.

Ontology Learning Tasks
Terms
- Synonyms
- Concepts
- Concept Hierarchies
- Relations
- Rules

Learning Relations in SKOS Model
- "Broader" (approximately equivalent to subsumption) and "related".
- "Information Theory Principle for Concept Relationship"

Information Theory Principle for Concept Relationship: A concept Cp is broader than another concept Cq if the following two conditions hold:
1. (Simplicity condition) the similarity measure between them is greater than a certain threshold, and
2. (Divergence difference condition) the difference between Kulback-Leibler divergence measures.
\[ D_{KL}(P||Q) - D_{KL}(Q||P) < 0 \]

Ontology Learning Algorithms
- Concepts are extracted from corpus and represented as documents using words in those documents that are annotated using the concepts.
- Computation of relations between concepts is transformed into computation of relations between representing documents.
- Computation is based on probabilistic topic models.

Datasets
- Crawler and scraper for web pages collected from digital libraries.
- Tokenising, stopwords removing, POS tagging, stemming and indexing.
- Concept extraction and representation.

Training pLSA and LDA
- pLSA uses Expectation-Maximisation algorithm.
- LDA uses Gibbs sampling.
- 30-90 topics are experimented for training.
- Concepts represented using documents are folded in learned topic models using same algorithms.

Experiment and Evaluation
- 672 sets of ontology statements are generated and evaluated by domain experts.
- In almost all of the cases precision of ontology using LDA is better than pLSA.
- The best precision using LDA is 86.6% and the worst is 58%. The best precision using PLSA is 80% and the worst is 39%.
- The possible reason is the generalisability of LDA to new documents.

Wang Wei, Payam M. Barnaghi, Andrzej Bargiela
(eysoww.payam.barnaghi, Andrzej.Bargiela)@nottingham.edu.my