TMRA 2007

Automatic Topic Map Generation from Free Text using Linguistic Templates

Ann Houston **Grammarsmith**

MOTIVATION

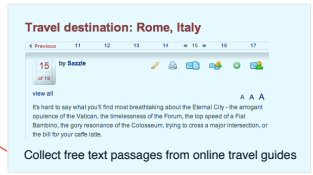
Free text contains vast amounts of useful information
Information extraction from free text is not easy
- Complexity of natural language
- Complexity of NLP technologies

Mapping is from 'messy' NL to well-defined TAOs

HYPOTHESIS

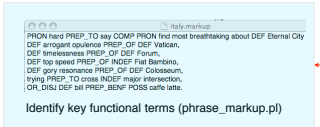
Can a set of linguistic templates extract TAOs automatically to form a useful topic map base from free text?
One pass pattern matching, no lexicon, no parse trees

APPROACH



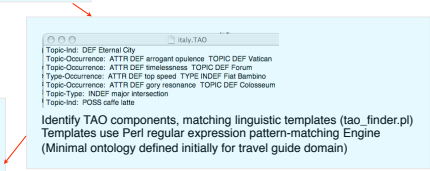
Collect free text passages from online travel guides

Split free text into major clauses (phrase_match.pl)



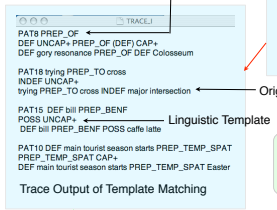
Identify key functional terms (phrase_markup.pl)

Generate xtm code and explore output in Omnigator



Identify TAO components, matching linguistic templates (tao_finder.pl)
Templates use Perl regular expression pattern-matching Engine
(Minimal ontology defined initially for travel guide domain)

Template ID
and unmatched portion of phrase

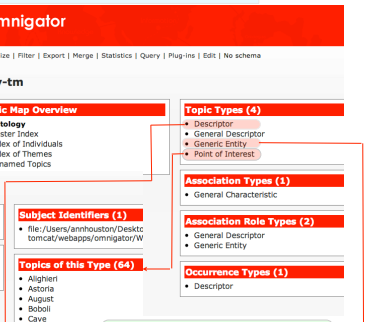


Original marked-up phrase

Linguistic Template

Trace Output of Template Matching

Rapid iterative cycle is possible through use of advanced visualization tool for Topic Maps, Ontopia's Omnigator



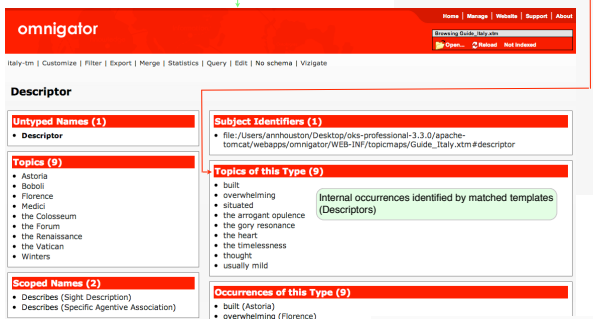
Individual Topics identified by matched templates (Point of Interest)

Template Misidentifications of Points of Interest

Tracing captures portions of phrases that are not consumed by the template matching.

These are important for debugging current templates and also determining what new templates should be added.

Could add 'Remaining Phrase' as Topic Type and inspect these via Omnigator as well.

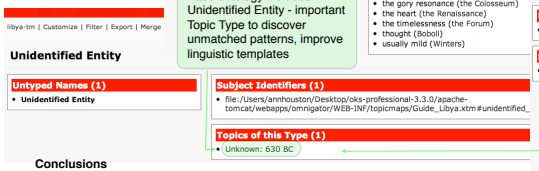


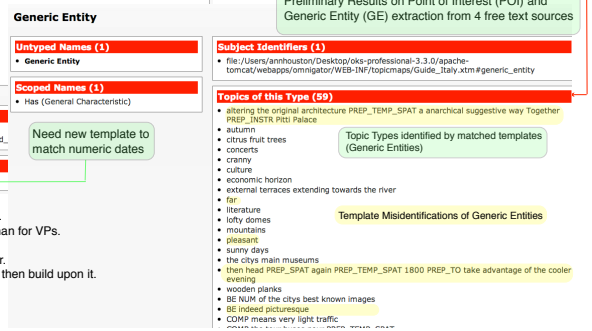
Internal occurrences identified by matched templates (Descriptors)

	Recall		Precision	
	POI	GE	POI	GE
Italy	64/84 0.76	59/77 0.77	55/64 0.86	40/59 0.68
Libya	116/162 0.72	104/150 0.69	101/115 0.88	68/104 0.65
Yellowstone	17/29 0.59	39/55 0.71	16/17 0.94	28/39 0.72
Chichen Itza	17/24 0.71	14/15 0.93	15/17 0.88	9/14 0.64

Preliminary Results on Point of Interest (POI) and Generic Entity (GE) extraction from 4 free text sources

Base ontology includes Unidentified Entity - important Topic Type to discover unmatched patterns, improve linguistic templates





Topic Types identified by matched templates (Generic Entities)

Need new template to match numeric dates

Template Misidentifications of Generic Entities

Conclusions

A simple set of linearly-applied 'shallow orthographic/functional' templates can extract TAOs.
Noun phrases contain a lot of the basic information, templates for NPs easier to formulate than for VPs.
Linguistic templates are language specific, and to some extent, domain specific.
Approach is fast, iterative and utilizes robust existing visualization tool - Ontopia's Omnigator.
Approach is a fast bootstrap to create topic maps from minimal ontology, human editing can then build upon it.
Tracing tools are very important in refining the developing set of linguistic templates.

Issues for Further Research

Ordering of Templates - implications for which templates will match phrases in which sequences
Coreference - recognize when 'the city' = 'Florence' (a difficult problem to solve automatically)
Conjunction Scope - distinguish 'expensive shops and restaurants' versus 'expensive shops and interesting restaurants'
Verbs - rich source of associations - probably need some lexicon to extract these, not just templates