

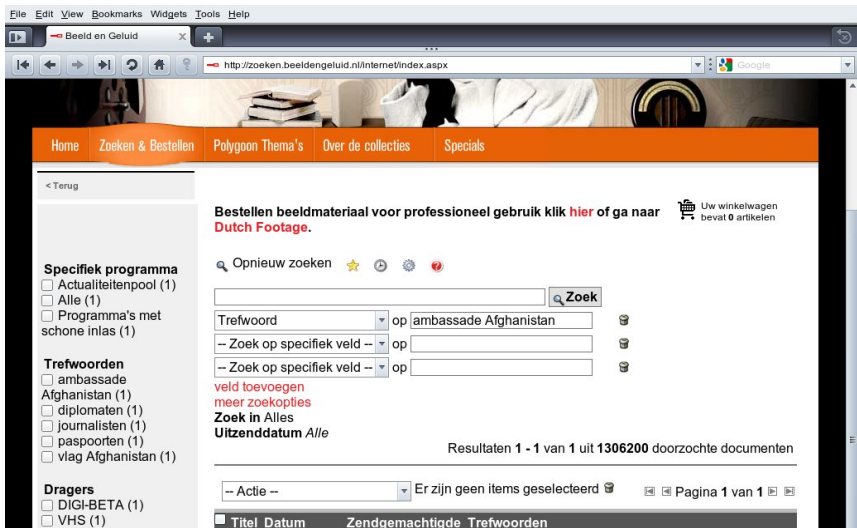
Cross-lingual Ontology Alignment using EuroWordNet and Wikipedia

Gosse Bouma

Information Science
University of Groningen

LREC 2010

Searching multimedia archive



The screenshot shows a web browser window with the address bar displaying `http://zoeken.beeldengeluid.nl/internet/index.aspx`. The website has an orange navigation bar with links: Home, Zoeken & Bestellen, Polygoon Thema's, Over de collecties, and Specials. On the left, there is a sidebar with filters for 'Specifiek programma' (Actualiteitenpool, Alle, Programma's met schone inlas), 'Trefwoorden' (ambassade Afghanistan, diplomaten, journalisten, paspoorten, vlag Afghanistan), and 'Dragers' (DIGI-BETA, VHS). The main content area features a search bar with the text 'Opnieuw zoeken' and a search button. Below the search bar, there are three rows of search criteria: 'Trefwoord' (set to 'ambassade Afghanistan'), and two rows for 'Zoek op specifiek veld'. A red text prompt 'veld toevoegen meer zoekopties' is visible. Below the search criteria, there is a section for 'Zoek in Alles' and 'Uitzenddatum Alle'. The results section shows 'Resultaten 1 - 1 van 1 uit 1306200 doorzochte documenten'. At the bottom, there is a table with columns: Titel, Datum, Zendgemachtigde, and Trefwoorden. The table is currently empty, and a message 'Er zijn geen items geselecteerd' is displayed.

File Edit View Bookmarks Widgets Tools Help

Beeld en Geluid

http://zoeken.beeldengeluid.nl/internet/index.aspx

Google

Home Zoeken & Bestellen Polygoon Thema's Over de collecties Specials

< Terug

Specifiek programma

- ☐ Actualiteitenpool (1)
- ☐ Alle (1)
- ☐ Programma's met schone inlas (1)

Trefwoorden

- ☐ ambassade Afghanistan (1)
- ☐ diplomaten (1)
- ☐ journalisten (1)
- ☐ paspoorten (1)
- ☐ vlag Afghanistan (1)

Dragers

- ☐ DIGI-BETA (1)
- ☐ VHS (1)

Bestellen beeldmateriaal voor professioneel gebruik klik **hier** of ga naar **Dutch Footage**.

Opnieuw zoeken

Zoek

Trefwoord op ambassade Afghanistan

-- Zoek op specifiek veld -- op

-- Zoek op specifiek veld -- op

veld toevoegen
meer zoekopties

Zoek in Alles

Uitzenddatum Alle

Resultaten 1 - 1 van 1 uit 1306200 doorzochte documenten

-- Actie -- Er zijn geen items geselecteerd

Pagina 1 van 1

Titel Datum Zendgemachtigde Trefwoorden

Multilingual Access

GTAA Thesaurus

- Dutch Institute for Sound and Vision
- archives tv and radio fragments
- Indexed by thesaurus keywords
- [subject](#) (3,887), [location](#) (13,992), [person](#) (97,617), [organizations and misc](#) (27,104)

Multilingual Access: How to support users not fluent in Dutch?

- Provide English keywords
 - Using terms from English [WordNet](#)
 - Using terms from English [Wikipedia](#)

Multilingual Access

GTAA Thesaurus

- Dutch Institute for Sound and Vision
- archives tv and radio fragments
- Indexed by thesaurus keywords
- [subject](#) (3,887), [location](#) (13,992), [person](#) (97,617), [organizations and misc](#) (27,104)

Multilingual Access: How to support users not fluent in Dutch?

- Provide English keywords
 - Using terms from English [WordNet](#)
 - Using terms from English [Wikipedia](#)

Cross-lingual Resources

From Dutch keywords to English Wordnet?

- **EuroWordNet**: links Dutch synsets to English synsets

From Dutch keywords to English Wikipedia?

- Dutch (& English) Wikipedia **cross-language links** and more...

Cross-lingual Resources

From Dutch keywords to English Wordnet?

- **EuroWordNet**: links Dutch synsets to English synsets

From Dutch keywords to English Wikipedia?

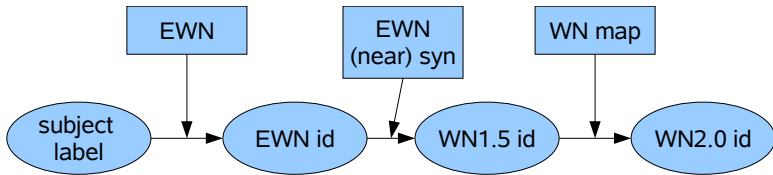
- Dutch (& English) Wikipedia **cross-language links** and more...

OAEI Workshops

Ontology Alignment Evaluation Initiative

- Organizes ontology alignment tasks and evaluation
- **Very Large Cross-lingual Resources**
 - One of the 2009 tasks
 - Aligning GTAA thesaurus with WN 2.0 and dbpedia
 - 2 participants in this task

GTAA → EWN → WN



- (Dutch) WordNet contains mostly subject terms, very few named entities
- Mapping from WN 1.5 to WN 2.0 : Daude, Prado, Rigau, *Mapping wordnet using structural information*, ACL 2000.

Linguistic Preprocessing

Stemming and Compound analysis

GTAA term	root	English
armoede	armoede	<i>poverty</i>
presidenten	president	<i>president</i>
biotechnologie	bio_technologie	<i>biotechnology</i>

Linking GTAA stems to EWN

- Both GTAA and EWN stemmed using Alpino (van Noord, 2009)
 - Ensures consistent analysis of compounds
- Compounds could be linked to more general concepts
 - *bio_technologie* **hyponym** of *technologie*
 - Providing hypernym relations was not part of the VLCR task

Linguistic Preprocessing

Stemming and Compound analysis

GTAA term	root	English
armoede	armoede	<i>poverty</i>
presidenten	president	<i>president</i>
biotechnologie	bio_technologie	<i>biotechnology</i>

Linking GTAA stems to EWN

- Both GTAA and EWN stemmed using Alpino (van Noord, 2009)
 - Ensures consistent analysis of compounds
- Compounds could be linked to more general concepts
 - *bio_technologie* **hyponym** of *technologie*
 - Providing hypernym relations was not part of the VLCR task

Linguistic Preprocessing

Stemming and Compound analysis

GTAA term	root	English
armoede	armoede	<i>poverty</i>
presidenten	president	<i>president</i>
biotechnologie	bio_technologie	<i>biotechnology</i>

Linking GTAA stems to EWN

- Both GTAA and EWN stemmed using Alpino (van Noord, 2009)
 - Ensures consistent analysis of compounds
- Compounds could be linked to more general concepts
 - *bio_technologie* **hyponym** of *technologie*
 - Providing hypernym relations was not part of the VLCR task

Aligning GTAA subject terms to EWN and WN

Coverage

subject labels	3,887	
linked to Dutch EWN	2,617	(67%)
unique ILIs	3,703	
avg. ambiguity	1.4	
linked to WN2.0	2,392	(62%)
unique synsets	3,676	
avg. ambiguity	1.5	

Ambiguity

- How to select the correct Dutch EWN synset?
- How to select the correct English WN 2.0 translation?

Aligning GTAA subject terms to EWN and WN

Coverage

subject labels	3,887	
linked to Dutch EWN	2,617	(67%)
unique ILIs	3,703	
avg. ambiguity	1.4	
linked to WN2.0	2,392	(62%)
unique synsets	3,676	
avg. ambiguity	1.5	

Ambiguity

- How to select the correct Dutch EWN synset?
- How to select the correct English WN 2.0 translation?

Word Sense Disambiguation

(Ambiguous) Mappings

Label	EWN	WN 2.0	WN Label
Ambassades	32482	103163291	embassy-noun-1
Ateliers	15723	104177492	studio_apartment-noun-1
Ateliers	15723	102651812	artist_s_workroom-noun-1
Schrijvers	19488	110090311	writer-noun-1
Schrijvers	25105	110090311	writer-noun-1
Anatomie	26879	104916889	human_body-noun-1
Anatomie	26879	105699031	anatomy-noun-1
Anatomie	32496	104916889	human_body-noun-1
Anatomie	32496	105699031	anatomy-noun-1

Word Sense Ambiguity

Wide-coverage method for Dutch?

- Find Predominant Word Senses (McCarthy et al., CL 2007)
- Used to align Dutch Wikipedia and EWN (Bouma, 2009)

concept	EWN synset	ILI	wn synset
Anatomie	↗ 26879	→ 105699031	→ anatomy-noun-1
	↘ 32496	→ 104916889	→ human_body-noun-2
	X		

No WSD...

- All WN senses are given as targets
- But in case of ambiguity with a lower confidence score

Word Sense Ambiguity

Wide-coverage method for Dutch?

- Find Predominant Word Senses (McCarthy et al., CL 2007)
- Used to align Dutch Wikipedia and EWN (Bouma, 2009)

concept	EWN synset	ILI	wn synset
Anatomie	↗ 26879	→ 105699031	→ anatomy-noun-1
	↘ 32496	→ 104916889	→ human_body-noun-2
		↗ ↘	

No WSD...

- All WN senses are given as targets
- But in case of ambiguity with a lower confidence score

Word Sense Ambiguity

Wide-coverage method for Dutch?

- Find Predominant Word Senses (McCarthy et al., CL 2007)
- Used to align Dutch Wikipedia and EWN (Bouma, 2009)

concept	EWN synset	ILI	wn synset
Anatomie	↗ 26879	→ 105699031	→ anatomy-noun-1
	↘ 32496	→ 104916889	→ human_body-noun-2
			

No WSD...

- All WN senses are given as targets
- But in case of ambiguity with a lower confidence score

OAEI Results

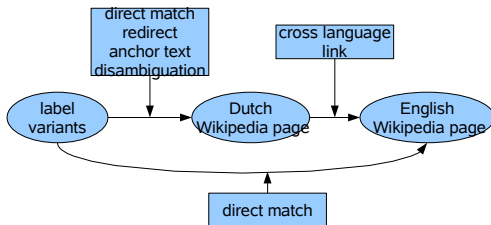
Official Evaluation Scores

	this paper			DSSim		
	# links	Prec	Rec	# links	Prec	Rec
Subject	3,663	0.59	0.59	655	0.77	0.19
Names	–	–	–	1,750	≈ 0.50	–

DSSim

- General purpose alignment tool
- Nagy, Vargas-Vera, Stolarski, 2009, *DSSim results for OAEI*

Aligning GTAA to Wikipedia (EN)



Dutch Wikipedia (2008) Statistics

page titles	656K
redirects	198K
unique anchors	1.1M
disambiguation pages	29K
cross-language links (NL→EN)	313K

Preprocessing

Generate Label Variants

- adelaars (*eagles*) → Adelaar
- adelaars → arenden (alternative GTAA label)
- Abbado, Claudio → Claudio Abbado

Examples

GTAA	match	Dutch Wikipedia	English Wikipedia
Aamodt, Kjetil Andre	redirect	Kjetil André Aamodt	Kjetil André Aamodt
Abbos, Samira	redirect	Samira Bouchibti	–
Abbado, Claudio	NL page	Claudio Abbado	Claudio Abbado
Abalkin, Leonid	EN page	–	Leonid Abalkin
Aleksej	anchor	Aleksej Nikolajevitsj van Rusland	Alexei Nikolaevich, Tsarevich of Russia
bedreigde diersoorten	redirect	Bedreigde soort	Endangered species
schrijvers (auteurs)	NL match	Auteur	Author
(<i>writers (authors)</i>)			
abortusklinieken	anchor	Abortus	Abortion
(<i>abortion clinics</i>)			
computerspelletjes	anchor	Videospel	Video game

Coverage

link type	subject		misc name		location		person	
	links	%	links	%	links	%	links	%
nl page	2,027	52.3	3,128	11.5	5,135	36.7	7,311	7.5
redirect	423	10.9	984	3.6	400	2.9	762	0.8
anchor	621	16.0	616	2.3	357	2.6	176	0.2
en page	260	6.7	4,085	15.1	3705	26.5	9,246	9.5
linked	3,127	80.6	8,830	32.6	9,602	68.6	17,521	17.9
no-english	357	9.2	2,197	8.1	878	6.3	5,721	5.9
no-link	394	10.2	16,077	59.3	3,512	25.1	74,375	76.2
total	3,878	100.0	27,104	100.0	13,992	100.0	97,617	100.0

Results

	this paper			DSSim		
	# links	Prec	Rec	# links	Prec	Rec
subject-dbp	3,381	0.86	0.62	1,363	0.70	0.30
person-dbp	17,516	0.91	–	2,238	0.79	–
misc-name-dbp	9,023	0.63	–	3,989	0.64	–
location-dbp	9,527	0.94	–	5,566	0.80	–

Discussion

Cross-lingual Resources

- Using cross-lingual information improves recall dramatically
 - 6x more recall than (generic) DSSim approach on linking concepts to WN
 - 3x more recall for linking terms to Wikipedia

Disambiguation Strategies

- *A4* → A4 (highway in the Netherlands), A4 (highway in Austria), A4 (paper format)
- Use categorical information in GTAA (*A4* is a location keyword)
- *Carole Lombard* → actress, ship
- Use scope note (*ship*) from GTAA
- Terms with same scope note map to pages in same category?

Discussion

Cross-lingual Resources

- Using cross-lingual information improves recall dramatically
 - 6x more recall than (generic) DSSim approach on linking concepts to WN
 - 3x more recall for linking terms to Wikipedia

Disambiguation Strategies

- *A4* → A4 (highway in the Netherlands), A4 (highway in Austria), A4 (paper format)
- Use categorical information in GTAA (*A4* is a location keyword)
- *Carole Lombard* → actress, ship
- Use scope note (*ship*) from GTAA
- Terms with same scope note map to pages in same category?

Discussion

Cross-language Links

- From EN $\rightarrow$ NL or NL $\rightarrow$ EN ?
- Cross-language links are not always reversible!
- EN $\rightarrow$ NL: *Bowling, Ten pin bowling* $\rightarrow$ *Bowling (NL)*
- NL $\rightarrow$ EN: *A4 highway* $\rightarrow$ *highways in the Netherlands*