

Building and Using Geospatial Ontology in the BioCaster Surveillance System

Son Doan¹, Quoc-Hung Ngo², Ai Kawazoe¹, and Nigel Collier¹

¹ National Institute of Informatics, Tokyo, Japan

² University of Information Technology, Vietnam National University (HCM), Vietnam

Motivation

- Disease and location names are important information in any reported news story.
- There is currently a lack of available geospatial ontology that covers all countries and major cities with the longitude/latitude information for visualization purposes.

Key idea

- Build a geospatial ontology from Wikipedia (<http://www.wikipedia.org>).

- Use advanced natural language processing techniques to detect events (including location names) within news stories.

Results

- A geospatial ontology with two administrative levels: 243 names in level 1 and 4,025 names in level 2 with their part-whole relationship and longitudes/latitudes.
- Geospatial ontology was integrated into BioCaster ontology and freely downloaded at <http://biocaster.nii.ac.jp/index.php?page=ontology>.
- Algorithms for detecting locations are implemented inside the Global Health Monitor, publicly available at <http://biocaster.nii.ac.jp>.

- We collected data from a 10-week period (Dec 20, 2007 – Feb 20, 2008):
 - 7,412 English news.
 - Covering 110 countries and 360 sub-countries.
 - News by continents: 58.00% Africa, 18.23% Asia, 11.37% South America, 5.30 % North America, 3.40% Middle East, 2.86% Europe and 0.34% Ocean.

- The system successfully detected ebola in Uganda (Bundibugyo, Kampala, Mbarara), yellow fever in Brazil (Goias, Sao Paulo), avian influenza in Indonesia (Jakarta, Banten), and cholera in Vietnam (Ha Noi, Ha Tay).

Challenges and Future work

- Geo-coding: To disambiguate location names, e.g., Camden can be a area in London (UK) or a town in New South Wales (Australia).
- Extend to deeper administrative levels like districts and sub-districts (wards, towns, villages).
- Evaluate and compare to other available resources, i.e., GAZ, dbpedia will be considered.

Acknowledgements

This work was supported by Grants-in-Aid from the Japan Society for the Promotion of Science (grant no. 18049071) and Transdisciplinary Integration Grant from ROIS.

Building a geospatial ontology from Wikipedia

Step 1 Extract names and ISO 3166-1 code of countries and dependent territories (level 1).

For each country, extract names and ISO 3166-2 codes of country subdivisions and dependent areas (level 2).

Step 2 Re-create the part-whole relations between level 1 and level 2 of geographical ontology.

Step 3 Verify results by manually checking local administrative websites.

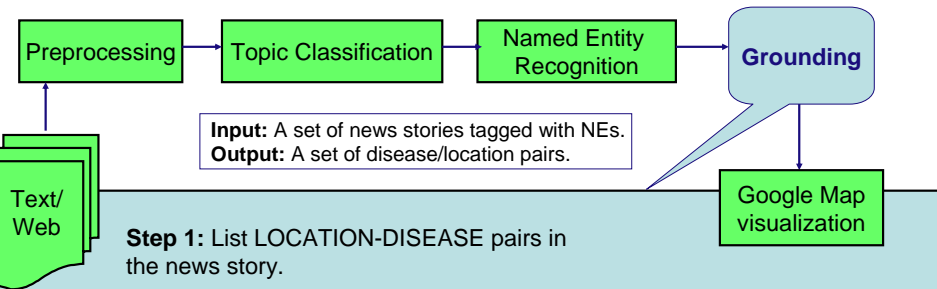
In addition we extracted geo-coordinates of all countries and sub-countries for visualization purpose.

http://en.wikipedia.org/wiki/ISO_3166-1

Official country names used by the ISO 3166 MA	Region	Alpha-3	Alpha-2	Local ISO codes
Algeria	AF	ALG	AF	ISO 3166-2:AF
Angola	AF	AGO	AG	ISO 3166-2:AG
Argentina	SA	ARG	SA	ISO 3166-2:SA
Armenia	AS	ARM	AS	ISO 3166-2:AS
Australia	OC	AUS	AU	ISO 3166-2:AU
Austria	EU	AUT	EU	ISO 3166-2:EU
Bahrain	AS	BHR	AS	ISO 3166-2:AS
Bangladesh	AS	BAN	AS	ISO 3166-2:AS
Barbados	NA	BRB	NA	ISO 3166-2:NA
Belarus	EU	BLR	EU	ISO 3166-2:EU
Belgium	EU	BEL	EU	ISO 3166-2:EU
Belize	NA	BZL	NA	ISO 3166-2:NA
Benin	AF	BEN	AF	ISO 3166-2:AF
Bhutan	AS	BUT	AS	ISO 3166-2:AS
Bolivia	SA	BOL	SA	ISO 3166-2:SA
Bosnia and Herzegovina	EU	BOS	EU	ISO 3166-2:EU
Brazil	SA	BRA	SA	ISO 3166-2:SA
Bulgaria	EU	BGR	EU	ISO 3166-2:EU
Burkina Faso	AF	BUR	AF	ISO 3166-2:AF
Burundi	AF	BUR	AF	ISO 3166-2:AF
Cambodia	AS	CMR	AS	ISO 3166-2:AS
Canada	NA	CAN	NA	ISO 3166-2:NA
Chad	AF	CHA	AF	ISO 3166-2:AF
Chile	SA	CHL	SA	ISO 3166-2:SA
China	AS	CHN	AS	ISO 3166-2:AS
Colombia	SA	COL	SA	ISO 3166-2:SA
Costa Rica	NA	CRI	NA	ISO 3166-2:NA
Cote d'Ivoire	AF	CIV	AF	ISO 3166-2:AF
Croatia	EU	HRV	EU	ISO 3166-2:EU
Cuba	NA	CUB	NA	ISO 3166-2:NA
Cyprus	EU	CYP	EU	ISO 3166-2:EU
Czechia	EU	CZE	EU	ISO 3166-2:EU
Denmark	EU	DNK	EU	ISO 3166-2:EU
Dominican Republic	NA	DOM	NA	ISO 3166-2:NA
Dominica	NA	DMA	NA	ISO 3166-2:NA
DRC	AF	COD	AF	ISO 3166-2:AF
Ecuador	SA	ECU	SA	ISO 3166-2:SA
Egypt	AF	EGY	AF	ISO 3166-2:AF
El Salvador	NA	SLV	NA	ISO 3166-2:NA
Equatorial Guinea	AF	GNQ	AF	ISO 3166-2:AF
Eritrea	AF	ERI	AF	ISO 3166-2:AF
Estonia	EU	EST	EU	ISO 3166-2:EU
Ethiopia	AF	ETH	AF	ISO 3166-2:AF
Finland	EU	FIN	EU	ISO 3166-2:EU
France	EU	FRA	EU	ISO 3166-2:EU
Germany	EU	DEU	EU	ISO 3166-2:EU
Ghana	AF	GHA	AF	ISO 3166-2:AF
Greece	EU	GRC	EU	ISO 3166-2:EU
Greenland	NA	GRL	NA	ISO 3166-2:NA
Guatemala	NA	GTM	NA	ISO 3166-2:NA
Guinea	AF	GIN	AF	ISO 3166-2:AF
Guinea-Bissau	AF	GNB	AF	ISO 3166-2:AF
Haiti	NA	HTI	NA	ISO 3166-2:NA
Honduras	NA	HON	NA	ISO 3166-2:NA
Hungary	EU	HUN	EU	ISO 3166-2:EU
India	AS	IND	AS	ISO 3166-2:AS
Indonesia	AS	IDN	AS	ISO 3166-2:AS
Iran	AS	IRN	AS	ISO 3166-2:AS
Iraq	AS	IRQ	AS	ISO 3166-2:AS
Ireland	EU	IRL	EU	ISO 3166-2:EU
Israel	AS	ISR	AS	ISO 3166-2:AS
Italy	EU	ITA	EU	ISO 3166-2:EU
Jamaica	NA	JAM	NA	ISO 3166-2:NA
Japan	AS	JPN	AS	ISO 3166-2:AS
Jordan	AS	JOR	AS	ISO 3166-2:AS
Kazakhstan	AS	KAZ	AS	ISO 3166-2:AS
Kenya	AF	KEN	AF	ISO 3166-2:AF
Korea	AS	KOR	AS	ISO 3166-2:AS
Kuwait	AS	KWT	AS	ISO 3166-2:AS
Kyrgyzstan	AS	KGZ	AS	ISO 3166-2:AS
Laos	AS	LAO	AS	ISO 3166-2:AS
Latvia	EU	LVA	EU	ISO 3166-2:EU
Lebanon	AS	LBN	AS	ISO 3166-2:AS
Lesotho	AF	LSO	AF	ISO 3166-2:AF
Lithuania	EU	LIT	EU	ISO 3166-2:EU
Luxembourg	EU	LUX	EU	ISO 3166-2:EU
Macao	AS	MCO	AS	ISO 3166-2:AS
Madagascar	AF	MAD	AF	ISO 3166-2:AF
Mali	AF	MLI	AF	ISO 3166-2:AF
Maldives	AS	MDV	AS	ISO 3166-2:AS
Mexico	NA	MEX	NA	ISO 3166-2:NA
Moldova	EU	MDA	EU	ISO 3166-2:EU
Mongolia	AS	MNG	AS	ISO 3166-2:AS
Montenegro	EU	MNE	EU	ISO 3166-2:EU
Morocco	AF	MAR	AF	ISO 3166-2:AF
Mozambique	AF	MUZ	AF	ISO 3166-2:AF
Myanmar	AS	MMR	AS	ISO 3166-2:AS
Nepal	AS	NPL	AS	ISO 3166-2:AS
Netherlands	EU	NLD	EU	ISO 3166-2:EU
New Zealand	OC	NZL	OC	ISO 3166-2:OC
Nicaragua	NA	NIC	NA	ISO 3166-2:NA
Niger	AF	NER	AF	ISO 3166-2:AF
Nigeria	AF	NGA	AF	ISO 3166-2:AF
North Macedonia	EU	MKD	EU	ISO 3166-2:EU
North Korea	AS	PRK	AS	ISO 3166-2:AS
Norway	EU	NOR	EU	ISO 3166-2:EU
Oman	AS	OMN	AS	ISO 3166-2:AS
Pakistan	AS	PAK	AS	ISO 3166-2:AS
Panama	NA	PAN	NA	ISO 3166-2:NA
Papua New Guinea	OC	PNG	OC	ISO 3166-2:OC
Paraguay	SA	PRY	SA	ISO 3166-2:SA
Peru	SA	PER	SA	ISO 3166-2:SA
Philippines	AS	PHL	AS	ISO 3166-2:AS
Poland	EU	POL	EU	ISO 3166-2:EU
Portugal	EU	PRT	EU	ISO 3166-2:EU
Romania	EU	ROU	EU	ISO 3166-2:EU
Russia	AS	RUS	AS	ISO 3166-2:AS
Rwanda	AF	RWA	AF	ISO 3166-2:AF
Saudi Arabia	AS	SAU	AS	ISO 3166-2:AS
Senegal	AF	SEN	AF	ISO 3166-2:AF
Serbia	EU	SRB	EU	ISO 3166-2:EU
Seychelles	AF	SYC	AF	ISO 3166-2:AF
Sierra Leone	AF	SLE	AF	ISO 3166-2:AF
Singapore	AS	SGP	AS	ISO 3166-2:AS
Slovakia	EU	SVK	EU	ISO 3166-2:EU
Slovenia	EU	SVN	EU	ISO 3166-2:EU
South Africa	AF	ZAF	AF	ISO 3166-2:AF
South Korea	AS	KOR	AS	ISO 3166-2:AS
Spain	EU	ESP	EU	ISO 3166-2:EU
Sri Lanka	AS	LKA	AS	ISO 3166-2:AS
Sweden	EU	SWE	EU	ISO 3166-2:EU
Switzerland	EU	CHE	EU	ISO 3166-2:EU
Taiwan	AS	TWN	AS	ISO 3166-2:AS
Tanzania	AF	TZA	AF	ISO 3166-2:AF
Togo	AF	TGO	AF	ISO 3166-2:AF
Tonga	OC	TON	OC	ISO 3166-2:OC
Turkey	AS	TUR	AS	ISO 3166-2:AS
Turkmenistan	AS	TKM	AS	ISO 3166-2:AS
Uganda	AF	UGA	AF	ISO 3166-2:AF
Ukraine	EU	UKR	EU	ISO 3166-2:EU
United Kingdom	EU	GBR	EU	ISO 3166-2:EU
United States	NA	USA	NA	ISO 3166-2:NA
Uruguay	SA	URY	SA	ISO 3166-2:SA
Uzbekistan	AS	UZB	AS	ISO 3166-2:AS
Venezuela	SA	VEN	SA	ISO 3166-2:SA
Vietnam	AS	VNM	AS	ISO 3166-2:AS
Yemen	AS	YEM	AS	ISO 3166-2:AS
Zambia	AF	ZMB	AF	ISO 3166-2:AF
Zimbabwe	AF	ZWE	AF	ISO 3166-2:AF

http://en.wikipedia.org/wiki/ISO_3166-2:AF

Algorithm for grounding disease/location into BioCaster ontology



- Step 1:** List LOCATION-DISEASE pairs in the news story.
- Step 2:** Calculate the frequency of LOCATION-DISEASE pairs in the news database for the last 12 hours.
- Step 3:** Rank LOCATION – DISEASE pairs by the frequencies calculated in Step 2. Use a threshold to choose top LOCATION - DISEASE names.
- Step 4:** Map disease and location names:
 - If sequences within DISEASE matches (regular expression matching) to a synonym in BioCaster ontology then DISEASE was assigned to that disease name.
 - If sequences within LOCATION matches (regular expression matching) to a location in geospatial ontology then LOCATION was assigned to that location name (countries and cities).
- Step 5:** Re-map into news stories: Match detected diseases and locations within the first half of each news story. If both disease and location are matched then they are stored; otherwise, skip.

Nature Precedings : doi:10.1038/npre.2008.2110.1 : Posted 22 Jul 2008