I WORKSHOP DE PROCESAMIENTO AUTOMATIZADO DE TEXTOS Y CORPORA (WoPATeC-2012)

Panel de Convergencia

Patrocinan

Motivación

- En este Worskhop nos proponemos integrar dos ámbitos disciplinares, a veces considerados opuestos: las ciencias del lenguaje y las ciencias del la computación: Objeto Texto / Procesamiento del Lenguaje Natural.
- Es habitual, por una parte, percibir cierto escepticismo entre algunos especialistas cuando se presentan resultados en lingüística, obtenidos o acreditados por medio de procedimientos matemáticos-computacionales.
- La sensación de que en los resultados en el ámbito del procesamiento computacional faltara mayor reflexión lingüística.
- Convocar a especialistas de ambos ámbitos para discutir acerca de métodos y técnicas, utilizados en el ámbito del procesamiento automatizado del lenguaje natural.

Sabela Fernández

- Grado Máximo e Institución donde lo obtuvo:
- Cargo e Institución actual:
- Línea(s) de Investigación:
- Proyecto(s) Actuales:
- Publicaciones más relevantes:
- Datos de contacto

Daniel Campos

- Grado Máximo e Institución donde lo obtuvo:
- Cargo e Institución actual:
- Línea(s) de Investigación:
- Proyecto(s) Actuales:
- Publicaciones más relevantes:
- Datos de contacto

César Aguilar

- Doctor en Lingüística, UNAM (México)
- Facultad de Letras, Pontificia Universidad Católica de Chile
- Ingeniería lingüística, semántica formal, gramática formal, desarrollo de ontologías
- Proyecto(s) Actuales: *Building a Generative Lexicon Model for Medical Documents in English and Spanish*
- Publicaciones más relevantes: Sierra G., Alarcón R., Aguilar C. y Bach C. (2010): "Definitional verbal patterns for semantic relation extraction". En Auger, A. & Barrière, C. (eds.), *Probing Semantic Relations: Exploration and Identification in Specialized Texts,* John Benjamins Publishing, Amsterdam/Philadelphia: 73-96.
- Acosta, O., Sierra, G. y Aguilar, C. (2011): "Extraction of Definitional Contexts using Lexical Relations", *International Journal of Computer Applications*, 34(6): 46-53.
- Ortega, R., Aguilar, C., Villaseñor, L., Montes, M. y Sierra, G. (2011): "Hacia la identificación de relaciones de hiponimia/hiperonimia en Internet", *Revista Signos. Estudios de Lingüística*, 44(75): 68-84.
- Datos de contacto:
  Tel. (+562)3547845; Email: Cesar.Aguilar72@gmail.com

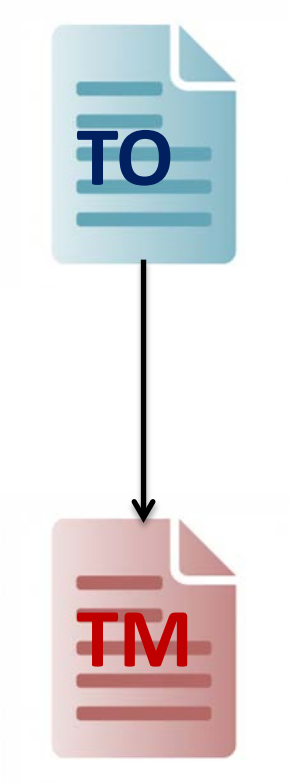
Rodrigo Alfaro

- Doctor(c) en Ingeniería Informática de la UTFSM.
- Académico de la Escuela de Ingeniería Informática de la PUCV.
- Director de Analitic S.A.
- Línea(s) de Investigación:
  - Clasificación Automática de Textos Multi-etiquetados
  - Clasificación Automática de Microblogs
- Proyecto(s) Actuales:
  - Nuevos modelos para Clasificación Automática de Textos Multi-etiquetados
  - Emprendimientos varios
- Publicaciones más relevantes:
  - "Text Representation in Multi-label Classification: Two New Input Representations". Alfaro R. & Allende H., presentado en la 10th International Conference on Adaptive and Natural Computing Algorithms (ICANNGA'11), Ljubljana, Slovenia. Artículo publicado en Lecture Notes in Computer Science, 2011, Volume 6594/2011, 61-70, DOI: 10.1007/978-3-642-20267-4_7
- Datos de contacto:
  Tel. (+56 32)2273609; Email: rodrigo.alfaro@ucv.cl; @rdgoalfaro; www.rodrigoalfaro.cl

Preguntas

1.- En el ámbito de las ciencias del lenguaje, ¿cuál ha sido el aporte de las técnicas y métodos de las ciencias computacionales en el estudio del lenguaje y en el desarrollo de aplicaciones lingüísticas? Dra. Sabela Fernández

Gestión de corpus y terminología en la traducción especializada



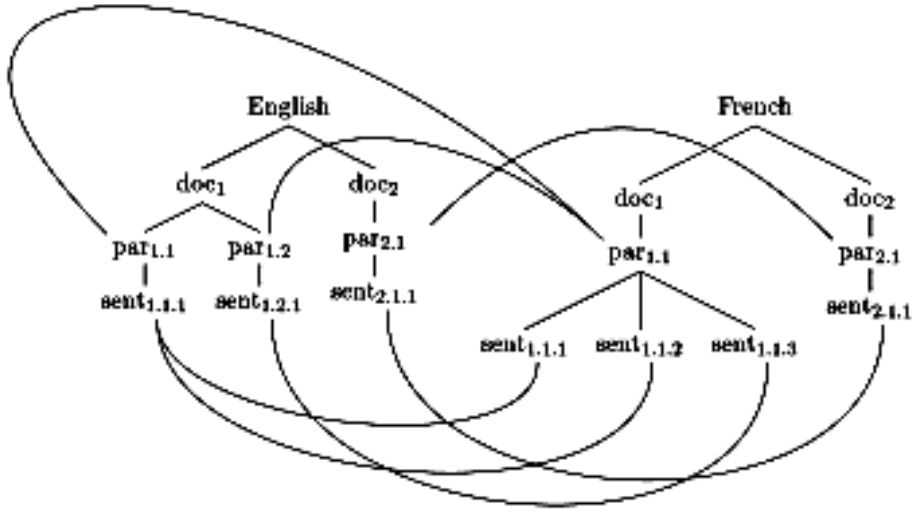
- Competencias cognitivas
- Competencias comunicativas
- Competencias lingüísticas

- Corpus
  - Corpus comparables
  - Corpus paralelos
- Terminología
  - N-gramas
  - Extracción de términos

Corpus

- Comparables
  - A partir de los motores de búsqueda en Internet
- Corpus paralelos
  - Alineación a distintos niveles

Documento
Párrafo
Frase
Palabra



Gestión de terminología

– N-gramas

bi-grama	frec. abs.	frec. rel.	bi-grama	frec. abs.	frec. rel.
da pesca	245	0,00066	productos do	19	0,00005
mollusques bivalves	129	0,00035	cabo fisterra	19	0,00005
do sector	127	0,00034	amoco cadiz	19	0,00005
na zona	121	0,00033	do plan	19	0,00005
bivalves vivants	116	0,00031	polo ige	19	0,00005
elaboración propia	115	0,00031	sobre todo	19	0,00005
da zona	107	0,00029	affaires maritimes	19	0,00005
da produción	106	0,00029	peso relativo	19	0,00005

– Extracción de terminología

- Estrategias lingüísticas (patrones sintácticos, recursos léxicos, recursos semánticos)
- Estrategias estadísticas (asociación)

Conclusión

- Colaboración muy fructífera
- Mejorar precisión en algunos resultados
- Mayor colaboración interdisciplinar
- Formación en estadística y computación en las ciencias sociales