Zirkuitu ibilbidea 1 - 1. pabilioia, 20160 LASARTE-ORIA Tel.: 943 37 18 49 Fax: 943 37 27 72
Balmes, 205 - 5º 2ª b 08006 BARCELONA Tel.: 93 415 61 31 Fax: 93 415 37 75
Xalbardin Baita 64122 URRUGNE (Hendaya) Tel. 33559 47 60 00 Fax 33559 47 60 01
http://www.ametza.com ametza@ametza.com

# PROYECTO JAZO 2007

## Título

Detección y seguimiento de sucesos para euskera y español

## Participantes

- Ametzagaiña A.I.E.

## Datos Generales

**Tipo:** Proyecto de Plan de Especialización
**Años de actividad:** 2007-2008

# Objetivos generales del proyecto

La Detección y Seguimiento de Sucesos (TDT, Topic Detection and Tracking) tiene como objeto el desarrollo de herramientas computacionales para el análisis de grandes colecciones de noticias, con el fin de extraer toda la información posible respecto a los sucesos por ellas narrados.

Es un campo de investigación que se engloba dentro del ámbito de los Sistemas de Extracción y Recuperación de Información, y que guarda estrecha relación con estos otros:

- Recuperación de Información (IR), indexación y búsqueda basada en texto libre.
- Extracción de Información (EI), elaboración automática de resúmenes o sumarización.
- Búsqueda de Respuestas (BR).

En anteriores ejercicios, nuestro centro de investigación ha realizado proyectos de especialización en los campos anteriormente citados, y el que proponemos ahora viene a completar una visión global y un tratamiento suficientemente exhaustivo del ámbito general de la Extracción y Recuperación de Información.

En nuestro caso, el punto de partida consta de varios flujos de información (diarios de información general) en dos idiomas (euskera y español).

En dichos flujos de información, las noticias suelen tratar de temas genéricos, que son asimilables a las categorías clásicamente utilizadas en clasificación de documentos (p.e. «economía», «deportes», etc). En estos temas, cada documento trata de un suceso de la actualidad (p.e. «la tasa de inflación de los últimos meses», «la final de la Copa América», etc.). Nos interesamos por la detección automática de estos sucesos y por el seguimiento automático de la información que les concierne.

Los sucesos no se conocen, por supuesto, a priori, pueden ocurrir en cualquier momento y su «vida» pueden ser muy variable. En términos de tratamiento automático, se trata de un problema de aprendizaje no supervisado dónde la evolución temporal del suceso desempeña un papel importante.

Las técnicas que se pretenden desarrollar deben ser genéricas, ya que deben poder utilizarse sobre numerosas fuentes de información diferentes, sin que eso dé lugar a adaptaciones costosas. Los datos normalmente implican información múltiple, los documentos de un mismo flujo de información pueden ser muy heterogéneos a la vez en su forma y su contenido.

Uno de los principales puntos de bloqueo a los que han llegado las investigaciones de los últimos años en este campo, estriba en la propia definición de lo que es un *suceso*. Interfieren discusiones que casi se podrían considerar filosóficas, a la

hora de distinguir entre tema *(topic)*, suceso *(event)*, actividad o dominio *(activity)*, ocurrencia *(occurrence)*... Dejando a un lado la idoneidad o no de cualquiera de las posibles opciones, es conveniente establecer a qué nos referimos en el presente proyecto cuando hablamos de suceso:

- **Suceso:** Acción, hecho o actividad que transcurre en un tiempo y lugar específicos. Por ejemplo: «Conciertos contra el cambio climático en julio de 2007».

Lo diferenciamos de:

- **Tema:** Concepto abstracto que agrupa a un conjunto de sucesos o actividades. Por ejemplo: «Cambio climático».
- **Actividad:** Un conjunto genérico conectado de acciones que transcurren en un tiempo y lugar específicos. Por ejemplo: «Campañas y medidas contra el cambio climático».
- **Ocurrencia:** Supone un nivel mayor de concreción en el tiempo o en el lugar que la considerada en el suceso. Por ejemplo: «El concierto de Australia dio inicio...». En nuestro proyecto, haremos referencia a éstas cuando tratemos de *sub-sucesos*.

## *OPORTUNIDAD DEL PROYECTO*

La información se ha convertido, más que nunca, en un factor de desarrollo económico, un motor de innovación. Las empresas pretenden mantener un conocimiento actualizado de su entorno global y de los dominios de su interés específico. La vigilancia tecnológica, económica, financiera, etc., es una de las herramientas esenciales de este conocimiento. La multiplicación de la cantidad de las fuentes de información impone desarrollar herramientas de tratamiento automático para dicha vigilancia. La vigilancia automático implica numerosas facetas.

Entre los problemas planteados por la vigilancia, nos interesamos aquí por el análisis en línea de flujo de información para que extraiga y que analice los sucesos que dependen de un conjunto de temas. En el ámbito de la informática, por ejemplo, un tema podrá ser los microprocesadores y un suceso la comercialización de un microprocesador particular en unas fechas concretas.

En el ámbito de la actualidad, existe numerosas fuentes emitiendo regularmente documentos datados. Es el caso por ejemplo de las agencias de prensa (Reuters, AFP, EFE...), los nuevos documentos llegan constantemente, pero es quizás en el mundo de los periódicos digitales donde el flujo de información se ha multiplicado de una forma más espectacular. Es tal la cantidad de datos y documentos que se ofrecen, que resulta imposible procesar siquiera una mínima parte de los mismos. El problema ya no es tanto tener acceso a mucha información, sino ser capaz de asimilarla y sintetizarla en base a unos intereses determinados de cada usuario.

# Análisis del estado del arte

Los sistemas para el seguimiento y detección de sucesos (TDT - Topic Detection and Tracking) comenzaron a partir de la iniciativa que se puso en marcha en 1997 por el DARPA (Gobierno de EEUU) dentro del programa TIDES. Dicha iniciativa convocó múltiples conferencias en los siguientes años. El propósito de dicha iniciativa era investigar nuevas técnicas computacionales para analizar las colecciones de noticias (habladas o escritas) y detectar los sucesos narrados por ellas.

A partir de aquellas primeras experiencias se ha creado un área de investigación dentro de la Recuperación de Información (RI), que hasta nuestros días a gozado del esfuerzo de numerosos grupos. He aquí las principales aproximaciones realizadas en este ámbito.

Una de las primeras pistas exploradas para la detección consistió en buscar cambios bruscos de vocabulario en el flujo de información. En particular se puede pensar que un nuevo suceso a menudo será caracterizado por una nueva serie de entidades nombradas. Las pruebas realizadas ponen de manifiesto que esta información es pertinente, pero ya desde un principio se vio que no conviene utilizarla sino en cooperación con otros métodos.

El enfoque mayoritario utiliza el modelo vectorial que en gran medida se emplea en RI. A partir de una codificación vectorial de los documentos y de una función de semejanza, se intenta detectar si un nuevo documento trata de un suceso ya presente en el flujo de información o introduce un nuevo suceso. Para eso se utiliza un límite máximo sobre la medida de semejanza. Se proponen dos grandes tipos de enfoque, uno se basa en algoritmos de clasificación incremental, y el otro en técnicas de «vecino más próximo». Esta última clase de métodos ofrece resultados sensiblemente mejores, si bien la primera tiene un coste menor, y permite realizar simultáneamente detección y seguimiento.

Un tercer enfoque está constituido por modelos probabilistas, como en otras tareas de RI; los modelos y los resultados son bastante similares a lo que se obtiene con los modelos vectoriales.

# Descripción de fases y tareas

## FASE 1: Procesamiento del lenguaje orientado a sucesos

Objetivo:

Aplicación de técnicas de PLN a los documentos para extraer toda aquella información significativa de cara al establecimiento de sucesos.

Descripción:

La teoría de la comunicación establece que una información de carácter noticioso debe cumplir la llamada «Teoría de las cinco W», es decir, debe dar respuesta a las cinco preguntas principales: dónde, cómo, quién, por qué y cuándo («where», «how», «who», «why» y «when», cinco partículas inglesas que bautizan dicha teoría).

Está claro que la pregunta más difícil de detectar con técnicas de PLN es la de «por qué», pero también es cierto que es la que menos información relevante aporta al objetivo del presente proyecto.

Del resto de las preguntas, el nivel de complejidad e importancia de cada una de ellas nos lleva a tratar por separado la información temporal («cuándo») del resto: «quién», «dónde» y «cómo».

Los tres últimos focos de información responden claramente al sujeto, a la localización y a otras circunstancias de la noticia: objeto, acción, cantidad...

El tratamiento de la información temporal se antoja básico en un sistema de seguimiento de sucesos, ya que las referencias y co-referencias temporales aportan numerosos datos para el establecimiento de relaciones entre los diferentes documentos que hacen referencia al mismo. El primer dato importante es la fecha de publicación, para a partir de ahí establecer las relaciones del documento respecto al hecho narrado, y entre los múltiples documentos referenciados ente sí.

Por último, no hay que olvidar que como resultado de nuestro proyecto LABUR-2005 —orientado a la sumarización— ya contamos con una representación del contenido de los documentos; es oportuno e interesante, por tanto, analizar la aportación que dichos resultados pueden realizar a la consecución del presente proyecto

## FASE 2: Agrupamiento de documentos orientado a sucesos

Objetivo:

Optimización y adaptación de las técnicas no supervisadas de agrupamiento de cara a la detección y seguimiento de sucesos.

Descripción:

El punto de partida de esta Fase es el resultado aportado por las técnicas de agrupamiento o clasificación no supervisada *(clustering)*¸ la cual, de una colección de documentos, establece grupos o *clusters*, en función de su semejanza formal. Dicho agrupamiento representa, en la práctica, una referencia bastante fiable de la similitud respecto a la proximidad del tema que tratan los diversos documentos.

Es necesaria una optimización de dicho mecanismo para su correcta utilización en el seguimiento de sucesos. Así, se ha de establecer cuáles son los métodos y las variables de agrupamiento más idóneas para tal fin. Dentro del tipo de agrupamiento, se probarán los algoritmos aglomerativos y los centroides. Para el primero, se experimentarán los diferentes parámetros (máximo, mínimo y porcentaje), y para el segundo, el número de iteraciones a realizar. Respecto a las variables, la principal es el umbral de semejanza, que puede ser variable dependiendo del objetivo de las diferentes tareas de la Fase 3 del proyecto.

Inicialmente se tratarán las noticias día a día, para posteriormente considerar períodos de mayor duración, principalmente a través de trigramas de días. En teoría, un suceso tiene un principio y un fin, aunque en la práctica dichos hitos son mucho más difusos. En esta tarea se trata de reunir el mayor número de datos que pueda aportar la herramienta de *clustering*, para su posterior proceso en la detección y seguimiento final de los sucesos.

Será necesario realizar reagrupamientos en función de la evolución temporal, tanto a futuro como en el pasado, combinando los datos (unión e intersección) de los

diferentes trigramas de días. Eso hará que haya que establecer mecanismos de inclusión y exclusión teniendo en cuenta la aparición de nuevos datos que pueden hacer que un documento se caiga o se incorpore.

## FASE 3: Detección y seguimiento de sucesos

Objetivo:

Identificación del suceso, del punto de inicio, y diseño de la representación y consulta del resultado.

Descripción:

Como parece lógico, esta es la Fase más compleja e importante del proyecto. En ella se identificará tanto el propio suceso, como las relaciones entre los documentos que le hacen referencia, y se planteará un sistema de consulta y recuperación de la información procesada.

De cara al establecimiento del suceso, es necesario, no solo definir el grupo de documentos que lo componen, sino también llegar a una formulación en lenguaje natural del «título» que definiría dicho suceso.

A menudo un suceso trae como consecuencia un hecho que con el tiempo se convierte asimismo en un suceso independiente. En vistas de la complejidad de decidir cuándo dicho suceso adquiere suficiente autonomía, se plantea establecer un sistema jerárquico de sucesos/sub-sucesos, que mantengan y representen la relación del descendiente con el antecesor.

A partir de la detección del suceso, y teniendo en cuenta el resultado de las Fases 1 y 2, se determinará la relación entre los diferentes documentos asignados al mismo. Dichas relaciones podrán ser sincrónicas (noticias del mismo día sobre el mismo hecho), o diacrónicas (evolución del tratamiento a través del tiempo). También hay que tener en cuenta que un documento puede hacer referencia a más de un suceso, por lo que podrá ser necesario asignarlos a varios grupos.

La detección del punto de inicio es un tema clave en la detección y seguimiento de sucesos, y dentro del cronograma está situado en un punto que puede parecer paradójico, pero que no es tal. Decidir cuál es la primera noticia sobre un suceso puede llevar a reconsiderar decisiones anteriores respecto a qué es un suceso y qué no. El sistema crea un nuevo candidato a suceso cuando un documento no se asemeja lo suficiente con ninguno de los registrados anteriormente, pero al considerar los textos con una perspectiva de duración en el tiempo, y con la aparición continua de nuevos documentos, las decisiones a tomar por el sistema pueden llevar a la reconsideración *a posteriori* de sucesos establecidos previamente.

Para la representación del resultado, se utilizarán principalmente diagramas de flujo que ofrezcan una primera visión lo más gráfica posible de la evolución del suceso en dos ejes: tiempo y cantidad. A partir de los resúmenes más visuales, se podrá profundizar más en los resultados, a través de círculos concéntricos incrementales: ficha de los documentos (título, procedencia, fecha...), pasajes relevantes extraídos previamente, y artículo completo.

En nuestro centro de investigación no entendemos que el Sistema de Detección y Seguimiento de Sucesos sea un desarrollo aislado e independiente, sino que guarda estrecha relación con los desarrollos realizados previamente en Recuperación y

Extracción de Información. Por tanto, se analizarán las posibilidades de su interacción con otros sistemas de los que ya disponemos, tales como generación automática de resúmenes o sistema de búsquedas de respuestas.

## FASE 4: Evaluación del sistema

Objetivo:

Establecimiento de la metodología para la evaluación, así como la batería de sucesos a detectar y seguir.

Descripción:

Para la evaluación de los resultados, se utilizará el llamado *Coste de Detección*, método establecido en la conferencia DTD-2000 y que hoy en día está considerado como el que mejor refleja la idoneidad de las técnicas empleadas en detección y seguimiento de sucesos. Para establecer dicho Coste de Detección, se utilizan las siguientes medidas: Cobertura, Precisión, Falsas Alarmas, y Errores por Omisión.

El aspecto más costoso de esta Fase reside en la confección del corpus de evaluación, ya que los únicos disponibles manejan idiomas diferentes a los tratados en nuestro proyecto. Será necesario, por tanto, elaborar un base de datos propia, para poder evaluar correctamente los resultados. Para ello, se utilizarán los anuarios que anualmente publican diversos medios de comunicación y, a partir de ahí, elegir una treintena de sucesos importantes, para posteriormente asignar manualmente la totalidad de textos que en el corpus global tengan relación con dichos sucesos.