DOI: 10.26314/GARR-Conf19-proceedings-14

Costruzione di un vocabolario controllato per la terminologia sulla Cybersecurity

Claudia Lanza

Università della Calabria

Abstract. Questo articolo si propone di presentare la realizzazione di un primo prototipo di vocabolario di termini controllato, un thesaurus, in lingua italiana per la sicurezza informatica all'interno di un progetto in collaborazione con l'Istituto di Informatica e Telematica (IIT) del Consiglio Nazionale delle Ricerche (CNR), CyberLab – Osservatorio sulla Cybersecurity (OCS). L'obiettivo principale è stato quello di fornire una risorsa in grado di comprendere quanta più informazione autorevole sul mondo della sicurezza informatica nonché offrire una piattaforma di condivisione per far meglio comprendere le interconnessioni terminologiche proprie di questo lessico specialistico sia agli utenti esperti che comuni.

Keywords. Cybersecurity; Thesaurus; Gestione della conoscenza; Formazione; Condivisione e riuso dei Dati.

Introduzione

Quando ci si confronta con domini che presentano un lessico specializzato, spesso risulta difficile comprendere pienamente il modo in cui i suoi concetti sono interconnessi tra di loro. È in questi casi che l'utilizzo di risorse di gestione e rappresentazione della conoscenza (*SKOS*) può fornire un importante contribuito alla comprensione di un lessico tecnico. Tra le risorse semantiche impiegate ai fini di un orientamento più facilitato alla conoscenza di ambiti specialistici, e alla condivisione dei dati, c'è il thesaurus. Come Broughton sottolinea, il thesaurus è spesso utilizzato sia per la rappresentazione dei termini più significativi associati a un dominio di studio che per l'indicizzazione dei documenti ad esso relativi.

La prima parte di questo articolo sarà dedicata alla presentazione della metodologia utilizzata per il recupero dei testi riferiti al mondo della Cybersecurity, di fatto la costruzione del corpus. La seconda parte si concentrerà sulla collaborazione con il gruppo di esperti di dominio, il gruppo dell'IIT presso il CNR di Pisa, con i quali si è avviato un progetto cooperativo per la costruzione di una piattaforma web rivolta a esperti del settore e a utenti comuni ai fini dello sviluppo di una risorsa online sulla sicurezza informatica; seguirà una breve descrizione della strategia di mappatura semantica tra i termini candidati per far parte del thesaurus italiano sulla sicurezza informatica e le tassonomie presenti negli standard di riferimento per la comunità degli esperti in Cybersecurity, i.e., NIST e ISO 27000:2016.

Infine l'articolo si concluderà con la delineazione della costruzione del thesaurus come guida e formazione per l'utente alla comprensione di tecnicismi.

1. Costruzione del corpus

Per la costruzione del corpus è stato seguito il sistema della gerarchia delle fonti secondo cui è preferibile organizzare il reperimento dei documenti sul settore scientifico di interesse partendo prima dal livello istituzionale-normativo. Il corpus ha iniziato ad essere popolato di decreti legislativi, norme, direttive ministeriali, regolamenti, tutti riferiti al campo della sicurezza informatica in lingua italiana. Successivamente, il bacino è stato allargato anche alle riviste di settore che hanno contribuito ad espandere sempre di più la terminologia di dominio. Infatti, la Cybersecurity presenta un lessico molto variegato, caratterizzato da un'importante impronta legislativa che ne gestisce la regolamentazione, nonché da tecnicismi adoperati da esperti di settore su riviste di dominio. Due delle principali fonti italiane divulgative analizzate sono le riviste "GNOSIS" e "Cybersecurity Trends", e, tra le linee guida, quelle pubblicate dal Computer Emergency Response Team (CERT), così come glossari specifici, e.g. "Glossario Intelligence" e "Threathsaurus".

2. Affiancamento con gli esperti di dominio

I documenti costituenti del corpus di partenza sono stati sottoposti a un processo di trattamento linguistico per estrarre i termini più rappresentativi che sarebbero diventati parte del thesaurus italiano sulla sicurezza informatica. Il software scelto per l'estrazione terminologica è Text To Knowledge (T2K) che permette di ottenere in output una lista di termini significativi in base alla formula Term Frequency–Inverse Document Frequency (TF-IDF): vengono ordinati in modo decrescente i termini considerati più significativi all'interno dei documenti dati in input come corpus di partenza.

A partire da questo elenco iniziale di termini candidati si è avviata una prima forma di collaborazione con gli esperti di dominio. Come Schultz evidenzia, per il lavoro del terminologo diventa molto importante la fase di affiancamento con il gruppo di esperti di dominio in quanto sono proprio questi ultimi a garantire un livello di valutazione appropriata circa la base contenutistica ottenuta dai software di estrazione semantica. I termini sono la rappresentazione linguistica dei concetti propri di un dominio specifico, e il thesaurus ha l'obiettivo di fornire una sistematica rappresentazione della struttura di conoscenza propria a determinati domini specialistici. Pertanto, la figura degli esperti diviene fondamentale ai fini della selezione più accurata dei termini che rappresentano il dominio. Con il gruppo di esperti sono stati definiti sia la determinazione di rapporti semantici esistenti tra i termini estratti dai documenti del corpus che l'inserimento di nuovi contenuti ai fini dell'arricchimento tesaurale.

2.1 Comparazione con gli standard di riferimento

Per ottenere un alto grado di accuratezza e precisione terminologica nella struttura tesaurale italiana per la sicurezza informatica, è stato avviato a monte un processo di mappatura tra le tassonomie presenti negli standard ufficiali di riferimento per gli esperti di Cybersecurity e le liste ottenute dall'estrazione terminologica con T2K. Le due banche dati ufficiali con cui sono stati mappati i termini derivati dal corpus di partenza sono il vocabolario e il glossario contenuti rispettivamente negli standard ISO/IEC 27000:2016

e NIST 7298. È stata verificata la presenza dei termini interni ai primi due elenchi sulla lista dei termini ottenuti con T2K per convalidare la copertura semantica data dai termini rappresentativi candidati e filtrati insieme agli esperti.

3. Conclusioni

L'obiettivo principale alla base dello sviluppo del thesaurus italiano per la Cybersecurity è stato quello di fornire uno tipo di supporto linguistico non ancora, secondo ricerche effettuate e sulla base delle nostre conoscenze, esistente ufficialmente nell'ambito italiano. Il thesaurus è strutturato sulla base di relazioni semantiche di base che gestiscono i rapporti tra i termini: la relazione di gerarchia che regola i rapporti tra i termini più generici (BT) e quelli più specifici (NT); la relazione di sinonimia tra i termini preferiti (USE) e quelli non preferiti (UF) e la relazione di associazione (RT). Questa struttura di gestione della conoscenza permette una visualizzazione dinamica dei termini costitutivi di un dominio tecnico, come quello della Cybersecurity. A fini esemplificativi si veda la Fig. 1 in cui il termine "Phishing" viene rappresentato in base a una serie di rapporti semantici che intrattiene con altri termini interni al thesaurus: "Phishing" è un termine più specifico (NT) di "Spam" (BT), che, a sua volta, è un termine più specifico di "Meccanismi di attacchi informatici".

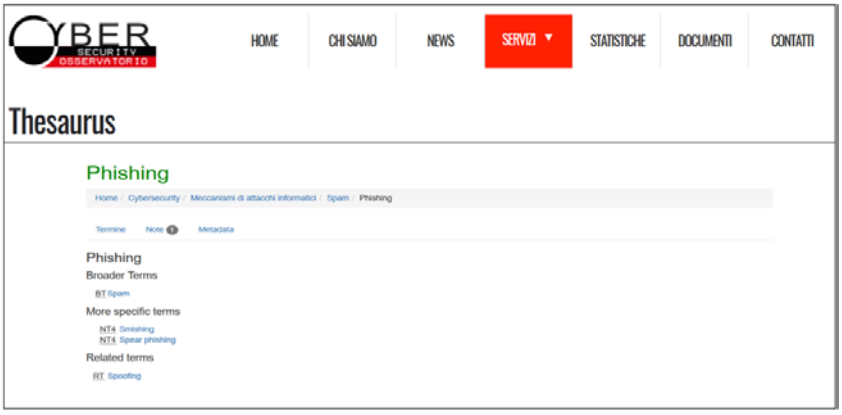


Fig. 1
"Phishing",
Thesaurus italiano
per la Cybersecurity

Riferimenti bibliografici

A. Milers, S. Bechoofer, "SKOS Simple Knowledge Organization SystemReference", W3C Recommendation 18 August 2009.

C. Sammut and G. I. Webb, "Encyclopedia of Machine Learning" (1st ed.). Springer Publishing Company, Incorporated, 2011.

CERT Nazionale Italia, < https://www.certnazionale.it/category/linee-guida/> (ultimo accesso 10/04/2019).

Claire K. Shultz, Wallace L. Schultz, and Richard H. Orr, "Evaluation of Indexing by Group Consensum" (Final Report, Contract No OEC 1-7-070622-3890), Bureau of Research Office of Education, U.S. Department of Health, Education and Welfare, August 30, 1968, p. 40.

Cybersecurity Osservatorio, <https://www.cybersecurityosservatorio.it/> (ultimo accesso 10/04/2019).

Cybersecurity Trends, < https://www.cybertrends.it/> (ultimo accesso 10/04/2019).

F. Dell'Orletta, G. V. (2014). "T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts", in Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014).

G. Zagrebelsky, "Il sistema costituzionale delle fonti del diritto", Torino, EGES; UTET, 1984, p. 67.

GNOSIS Rivista italiana di intelligence, <http://gnosis.aisi.gov.it/gnosis/Start.nsf/pages/home> (ultimo accesso 10/04/2019).

International Standard ISO/IEC 27000:2016 (E) Information technology – Security techniques – Information security management systems – Overview and vocabulary, Fourth edition 2016-02-05.

J. C. Sager, "A practical course in terminology processing", John Benjamins Publishing Company, 1990.

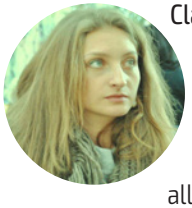
Presidenza del Consiglio dei Ministri – Sistema di informazione per la sicurezza della Repubblica, "Il linguaggio degli Organismi Informativi", Glossario Intelligence, <https://www.sicurezzanazionale.gov.it/sisr.nsf/quaderni-di-intelligence/glossario-intelligence.html> (ultimo accesso Ottobre 2018).

R. Kisserl, NISTIR 7298 National Institute of Standards and Technology Interagency or Internal Report 7298r2, 2013, May.

Sophos "Threatsaurus The A-Z of Computer and data security threats", <https://www.sophos.com/en-us/medialibrary/PDFs/other/sophosthreatsaurusaz.pdf?la=en> (ultimo accesso Settembre 2018).

W. Broughton, "Costruire Thesauri: strumenti per indicizzazione e metadati semantici" (a cura di) P. Cavaleri, (traduzione di) L. Ballestra e L. Venuti, Milano, EditriceBibliografica, 2008.

Autrice



Claudia Lanza - c.lanza@dimes.unical.it

Dottoranda in ICT presso l'Università della Calabria, Dipartimento DIMES, dal 2017. Ha compiuto gli studi specialistici a Milano presso l'Università Cattolica del Sacro Cuore in "Scienze del linguaggio, terminologie e tipologie dei testi" dove si è laureata nel 2016.

Attualmente è PhD Visiting fino al 2020 presso l'Université de Nantes, Laboratoire LS2N, all'interno del gruppo TALN (Traitement automatique du langage naturel). Il suo campo di ricerca comprende lo sviluppo di risorse per l'organizzazione e rappresentazione di lessici specialistici. In particolar modo, le aree di investigazione riguardano la costruzione di risorse semantiche come thesauri e ontologie e lo studio di metodi di linguistica computazionale per l'automatizzazione di operazioni semantiche.